Blind Image Deconvolution: Nonstationary Bayesian approaches to restoring blurred photos

Tom E. Bishop

A thesis submitted for the degree of Doctor of Philosophy.
The University of Edinburgh.
October 2008

Abstract

High quality digital images have become pervasive in modern scientific and everyday life — in areas from photography to astronomy, CCTV, microscopy, and medical imaging. However there are always limits to the quality of these images due to uncertainty and imprecision in the measurement systems. Modern signal processing methods offer the promise of overcoming some of these problems by post-processing these blurred and noisy images. In this thesis, novel methods using nonstationary statistical models are developed for the removal of blurs from out of focus and other types of degraded photographic images.

The work tackles the fundamental problem blind image deconvolution (BID); its goal is to restore a sharp image from a blurred observation when the blur itself is completely unknown. This is a "doubly ill-posed" problem — extreme lack of information must be countered by strong prior constraints about sensible types of solution. In this work, the hierarchical Bayesian methodology is used as a robust and versatile framework to impart the required prior knowledge.

The thesis is arranged in two parts. In the first part, the BID problem is reviewed, along with techniques and models for its solution. Observation models are developed, with an emphasis on photographic restoration, concluding with a discussion of how these are reduced to the common linear spatially-invariant (LSI) convolutional model. Classical methods for the solution of ill-posed problems are summarised to provide a foundation for the main theoretical ideas that will be used under the Bayesian framework. This is followed by an in-depth review and discussion of the various prior image and blur models appearing in the literature, and then their applications to solving the problem with both Bayesian and non-Bayesian techniques.

The second part covers novel restoration methods, making use of the theory presented in Part I. Firstly, two new nonstationary image models are presented. The first models local variance in the image, and the second extends this with locally adaptive non-causal autoregressive (AR) texture estimation and local mean components. These models allow for recovery of image details including edges and texture, whilst preserving smooth regions. Most existing methods do not model the boundary conditions correctly for deblurring of natural photographs, and a Chapter is devoted to exploring Bayesian solutions to this topic.

Due to the complexity of the models used and the problem itself, there are many challenges which must be overcome for tractable inference. Using the new models, three different inference strategies are investigated: firstly using the Bayesian maximum marginalised *a posteriori* (MMAP) method with deterministic optimisation; proceeding with the *stochastic* methods of variational Bayesian (VB) distribution approximation, and simulation of the posterior distribution using the Gibbs sampler. Of these, we find the Gibbs sampler to be the most effective way to deal with a variety of different types of unknown blurs. Along the way, details are given of the numerical strategies developed to give accurate results and to accelerate performance. Finally, the thesis demonstrates state-of-the-art results in blind restoration of synthetic and real degraded images, such as recovering details in out of focus photographs.

Declaration of originality

I hereby declare that, except as indicated below and in the text, the research described in this thesis and the thesis itself was composed and originated entirely by myself in the School of Engineering and Electronics at The University of Edinburgh. It has not been submitted for examination for any degree at any other University or as part of any professional qualification.

Chapters 4 and 5, and Appendix C contain some work that was published in the book chapter [25], of which I was the primary author. This work was completed during a study visit to the University of Granada during the second year of the PhD.

Chapter 7 contains results that were published in [24], and a preliminary version of the work in Chapter 9 was published in [26], while [27] forms the basis for part of Chapter 10.

Tom Bishop

Acknowledgements

Although this work has been a great personal endeavour over the past several years, there are many people who have been influential in helping me reach this point. I would like to begin by thanking my supervisor, James Hopgood, for his guidance, encouragement, and advice in this project — and of course his unique ability for tangential thinking that no doubt helped open my mind and formulate new ideas along the way.

During the second year of my PhD, I was fortunate enough to spend three and a half months working with Professor Rafael Molina at DECSAI in the Universidad de Granada, Spain. I wish to express my gratitude to Rafael for welcoming me during my stay, his insightful thoughts in our discussions, and for his continued advice and support. I should thank others who helped me out at DECSAI, not least Dácil Barreto and Francisco Palao, who stopped me running into too much trouble in the first months when my Spanish was dangerously poor, and who helped me to make the most of my visit! Thanks too to Derin Babacan for some interesting conversations during my last visit.

I am grateful to colleagues and friends at IDCOM who provided support, both in terms of discussions at the start of my PhD, and those who made the lab an interesting and lively place to be. In particular I would like to mention Peter Hillman, Christoph Burger-Scheidlin, Mark Sinclair, and Mika Levonen, and indeed all participating in those intense debates over lunch at Julie's Canteen early on (not to forget Julie's cuisine, without which I may have never have made it through). Luis, Elias, Yannis, Nedko, Fei, Athanasia, Cormac, and everyone from the Brass Monkey group also helped make sure we kept the right balance of work and fun! I wish to thank Christine Evers for helping proof read and assist with printing parts of the thesis at short notice! Of course many other friends, old and new, have provided encouragement and necessary distraction, and helped to shape my thoughts over these years. Joel Lindop no doubt played a role in influencing my ideas before I began the PhD, and my flatmates, in particular Alex Mantzaris, not only engaged me in varied academic and philosophical debate, but helped maintain a suitably crazy shared living environment that ensured there was never a dull moment out of the office! Finally I would like to thank my parents for accommodating me when other options ran out during the write up, and for all of their support over the years.

Contents

List of figures

Part I

Background, methods, and models

Chapter 1

Introduction and Problem Overview

Images are ubiquitous and indispensable in modern science and everyday life. Mirroring the abilities of our own human visual system, it is natural to display observations of the world in graphical form. Images are obtained in areas ranging from everyday photography to astronomy, remote sensing, medical imaging and microscopy. In each case, there is an underlying object or scene we wish to observe; the image is a visual representation of these observations.

Yet imaging, just as any other observation process, is never perfect: uncertainty creeps into the measurements, occurring as blur, noise, and other degradations in the recorded images. The image is a projection of the real world onto the lower-dimensional imaging medium, a procedure that intrinsically discards information. Sometimes the information lost may contain things we are interested in: it can be beneficial to try to recover these hidden details, to infer what the underlying scene that generated these observations really was.[1]

Recent signal processing techniques can provide a means to overcome some of the problems of the imperfect observation process, by post-processing these blurred and noisy images. By representing the observation process mathematically, and applying prior knowledge of the types of images we expect to see, restoration methods such as BID can be performed to recover detail and reduce image noise.

The rest of this chapter will consider the background to this problem, how and why we wish to restore blurred images, and several fields in which the techniques will be useful, as well as a brief history and outlook for the future. A more detailed survey will be given in the following chapters once the required formalisms for specifying the problem and solution methodologies have been introduced. This chapter concludes with an outline of the rest of the thesis.

[1] That there actually exists some fundamental exact representation of the world is an argument proposed by the Frequentist school of thought. As will be seen, the Bayesian approach extends the notion of the measurement process being responsible for the uncertainty to the concept that there is no one true reality, only a likely one that we can predict with a particular degree of uncertainty according to our observations and existing knowledge.

1.1 The Photographic Deblurring Problem

Blurry photographs are a frustratingly common sight to anyone who has tried taking pictures of friends on holiday or at a dimly lit social gathering. Despite recent advances in technology, users can often be disappointed that a significant proportion of these photos may end up being blurred; this can be due to motion of the subject, camera shake, or the subject being out of focus (see Fig. 1.1).

(a) A blurred photographic image (b) What we really want to see

Figure 1.1: *Blurred and sharp images: example representing motion blur due to camera shake*

The amount of available light restricts the clarity of the image, resulting in a trade-off between observation noise and sharpness: we can choose either a grainy but sharp image, or a smooth one but with motion blur; a longer exposure time permits more light to enter the camera giving a higher signal-to-noise ratio (SNR), but camera shake and movement in the scene blur details. This can pose a problem trying to image moving targets in dimly lit conditions. Using larger cameras and lenses can help, but often this is neither practical nor cost effective, and further physical limits such as diffraction and quantum noise are still met as resolution is pushed ever higher.

Of course, sometimes these images could be retaken in the hope that the next exposure will not be blurred, but frequently they are of some unique event that could only be captured once. It would be beneficial if these fleeting moments could be recovered and their details revealed.

1.2 Image Restoration and Blind Deconvolution

The field of digital image restoration is a signal processing discipline that attempts to restore the blurred and noisy observed image to give a better representation of the original scene. In doing so, the uncertainties in the observation process that have been described must be taken into account. The task of image restoration is termed an *ill-posed* inverse problem — this will be further discussed in §3.1. The key to being able to solve these types of problem is proper incorporation of prior knowledge about the original image and type of degradation into the restoration process.

Classical image restoration seeks an estimate of the true image assuming the blur is known. In contrast, *blind* image restoration tackles the much more difficult, but realistic, problem where the degradation is unknown. It is more difficult because there are a larger space of possible solutions: there are many blur and image combinations which could have resulted in something close to the observed image. The problem is to find reasonable ones that make sense according to some criteria. In general, the complete degradation in a observed blurred photograph is non-linear (including, for example, saturation, quantisation, compression) and spatially varying (non-uniform motion, imperfect optics); however in most work, it is assumed that the observed image is the output of a LSI system, representing convolution of the image and blur, to which noise is added. Therefore it becomes a BID problem, with the unknown blur represented as a PSF.

The BID problem represents one of the toughest challenges in image processing: the difficulty in correctly determining the "true" image that lies behind a blurry one is not inconsiderable; due to the fact that both the image and blur are unknown in the convolutive mixture, the degree of uncertainty is even higher, and the problem is doubly-ill-posed. The problem has received a considerable amount of research attention in recent years, yet no method has been proposed that can conclusively claim to have solved it.

Classical restoration has matured since its inception, in the context of space-exploration in the 1960's, where any attempt to digitally enhance the data was cheaper than the enormous

costs of re-imaging. Numerous techniques can be found in the literature (for recent reviews see [12, 106]). These differ primarily in the prior information about the image they include to perform the restoration task. The earliest algorithms to tackle the BID problem appeared as long ago as the mid 1970's [42, 208], attempting to identify known patterns in the blur; a small but dedicated effort followed through the late 80's (see for instance [7, 122, 126, 127, 210]), and a resurgence was seen in the 90's (see the reviews from this period in [115, 116]). Since then, the area has been extensively explored by the signal processing, astronomical, and optics communities. Many of the BID algorithms have their roots in estimation theory, linear algebra, and numerical analysis. As can be seen, the topic of BID spans a variety of areas and is very much a multi-disciplinary subject. This is also supported by the number and variety of areas in which the work may be applied, described in the following section.

1.3 Applications

The problem of photographic restoration has already been commented upon, and is one of the main motivations for the present work. In terms of the commercial potential, it is worth considering that the market for consumer digital photography has exploded over recent years, especially the camera phone industry which now accounts for the largest proportion of consumer imaging products sold. These products are driving a demand for miniaturisation and for higher quality pictures at reduced prices, which in turn puts a strain on hardware design, where cheap optics and small sensors must be used. In addition to the types of noise and blur already described that can plague small cameras, their cheaper optical elements often suffer from distortions and aberrations that introduce extra blur into the recorded images. Clearly to facilitate the desired miniaturisation, and overcome the problems of blurring, image restoration is a useful tool.

Aside from the application to improving the quality of everyday consumer photographs, there exist many other areas that can derive benefit from the application of BID techniques. Also concerning photographic images, there is much archive material sitting in picture libraries, news agencies, or private collections. Some of these images may pertain to newsworthy events, may represent items of historical interest, or merely hold sentimental value — faded old photographs of relatives for example. If they are partly or fully out-of-focus then restoration of these unique events will be valuable. In sporting events, a picture of an epic moment may have been captured, but it could be slightly mis-focussed. A somewhat controversial historical

example is the restoration of amateur footage from the assassination of the U.S. president John F Kennedy. Nowadays, it has become common for members of the public to send in pictures to news teams, who could derive value from enhancing these images if they are shot by an unsteady hand or do not show the subject clearly.

Along related lines, forensics is another key area where the recovery of details from blurred photographs is important. CCTV images are often inherently noisy and of low quality, captured in poorly lit areas. To correctly identify someone's face at the scene of a crime, it may be essential to make use of image restoration to restore the degraded image. Reading the letters on a motion blurred number-plate is another classic example where image deblurring can help.

Turning to more scientific applications, astronomical observations have been one of the fields that initiated and have most pushed the development of image restoration algorithms. The case of the Hubble space telescope (HST) is a good example, where spherical aberration defects in the main mirror of the multi-billion dollar orbiting observatory meant that the initial images (prior to an expensive repair) were not correctly formed. Consequently the image processing community was provided with an abundant supply of real-world data that astronomers wished to be restored, and this provided impetus for development of many restoration methods targeted for star-field images [114].

In both terrestrial based space-imaging, and remote sensing of the earth from satellites, the attainable resolution is usually limited by atmospheric effects. These effects include wavefront perturbation due to changes in refractive index through the turbid media, and light scattering from aerosols [97]. This results in a time-varying PSF at short durations, or a Gaussian-like blur when averaged over a longer exposure. When a series of short exposures with random PSFs are available, *multichannel* BID techniques are particularly appropriate [195].

Medical imaging is another field where image restoration has been essential for reconstruction of images from distorted projections. In fact, many medical images require the use of tomographic reconstruction techniques to view the data in the first place. This may be regarded though as just a different type of observation model. In images such as X-rays, and computed tomography (CT) scans, the degradations are unavoidable because of the low doses of radiation that can be used in order to protect the patients health. In modalities such as single photon emission computed tomography (SPECT), blurs may be due to radiation scattering and absorption within the patient, and collimation in the gamma camera [142]. Optical mi-

croscopy faces limits to resolving power, due for instance to diffraction. Although standard image restoration using PSF measurements taken using fluorescent beads is possible, these methods are not always accurate and blind restoration techniques have become popular in confocal microscopy. Ultrasound is another application where deconvolution methods can help restore images blurred, in this case, due to dispersive attenuation and aberrations in the tissue, as well as the finite spatial and frequency responses of the imaging system [141].

Related industrial applications that use reconstruction methods to image inside solid objects, such as non-destructive testing and geoscientific exploration using ground penetrating radar (GPR), can stand to benefit in a similar way from BID. Other important applications in science and industry include pre-processing of images for computer vision techniques; for example in optical character recognition (OCR), automatic face recognition, or the tracking of moving targets where motion blurring may be present [59].

In some cases the types of constraints available greatly simplify the problem: consider the OCR example; if the correct letter in the observed image can be recognised exactly from a database, there is no need to use complicated restoration techniques, as once recognised it may be reproduced to an arbitrary accuracy. However so-called "recogstruction" methods that fall between the two extremes, combining image restoration and recognition using a database of examples, have been also been proposed recently [10]).

Of course each of the fields described above have different types of requirements, and the algorithms that were originally developed for restoring astronomical or medical data may impose constraints such as availability of point sources in the image, or that the object has finite support — i.e. it is completely contained within the image, on a uniform black background. These constraints may not be so suitable for a more general class of photographic images, with which the work in this thesis is mainly concerned, hence the methods that will be considered need to be capable of modelling a more arbitrary range of images.

1.4 Philosophy of Restoration

An important question one may ask is "why restore" — why is BID useful? Could we not simply use a better observation procedure in the first place? Perhaps, but there always exist physical limits, such as photonic noise, diffraction or an observation channel outwith our control, and often images must be captured in sub-optimal conditions.

There are many existing images of unique events that cannot be retaken that we would like to be able to recover (for instance with forensics or archive footage); furthermore in these cases it is often infeasible to measure properties of the imaging system directly, because it is no longer available. Again this issue arises in medical imaging: it is preferable to avoid retaking a blurred x-ray, in order to safeguard the patient's health. Therefore in these cases, restoration procedures to enhance the existing images are the only feasible option to obtain a satisfactory image.

In terms of improving the observation process, there are hardware solutions to some of the problems in low-light photography: use of more expensive cameras, with larger lenses and sensors allows for faster shutter speeds and smaller apertures giving sharper images; optical image stabilisers can compensate to an extent for motion of the camera; but there are cases when even these can be ineffective, and they do not assist when the subject moves. In low-light photography, flash may be used to provide extra illumination to freeze motion, but built-in flash gives an unnatural and washed out appearance due to the location of the light source, and may be distracting or impossible to use in certain scenarios. Therefore to give natural looking photographs with compact cameras, some other solution is required.

Somewhat similarly to image stabilisation for regular cameras, adaptive optics using deformable mirrors are one solution to combat the problems of atmospheric turbulence in astronomy. These function by measuring and counteracting the distortions induced by the atmosphere in real-time. However they also have limitations, such as necessity of a guide star to act as a reference, and will not always be able to exactly cancel out the distortions. Therefore BID methods provide an alternative to improve image quality without requiring complicated real-time calibrations; alternatively BID may be used as a subsequent processing stage to further enhance the available resolving power.

Cost is another argument for using restoration for many applications. High quality optics, sensing equipment, and hardware corrections are expensive. However, processing power is abundant today and opens the door to the application of increasingly sophisticated models.

In the introduction, the nature of the imperfect imaging process was mentioned. The suggestion that signal processing offers a good way to gain the maximum useful information from the available signals is a sound one; however there exists an alternative argument that will be put forward further in the final chapter of this thesis. This is the suggestion that the restoration algorithm may become part of the imaging device itself, or that by designing both together

in a harmonious fashion, the types of measurements that we are really interested in obtaining can be acquired with greater expediency. This is a break from the traditional view that the measurement device is designed to gather the required information and signal processing is something of an afterthought to compensate for flaws in the measurement process. It is thought that this change in thinking could provide many advantages to the future of imaging.

1.5 Objectives of the Thesis

The central objective of the thesis will be to investigate algorithms for the blind restoration of blurred photographs[2]. An emphasis will be put on using *Bayesian* methods: as will be seen, these provide a robust and natural estimation framework for solving ill-posed problems, and allow prior information to be incorporated in a consistent manner. Furthermore, a key objective is to survey some of the diverse techniques that have been proposed previously for BID, and to illustrate the connections between them. The use of the Bayesian approach facilitates these comparisons in a straightforward way.

The *hierarchical* Bayesian paradigm will be used to develop the extended models that will be presented and investigated as part of the novel work in this thesis. This approach models the uncertainty in the model parameters as well as the observation model, and permits some of the difficulty in specifying their values to be transferred to providing a more detailed model.

To make use of the Bayesian framework, it will be necessary to develop appropriate models to represent the types of images and blurs that will be encountered in practice, and this will constitute one of the major investigative undertakings of the thesis. Statistical models will be developed in particular for natural images, although they are also capable of representing various types of blurs. These models are capable of preserving not only sharp edge transitions, as has been a common theme in may recent models, but also texture and fine detail in the restored images.

Finally, a number of different estimation procedures are possible for a given model under the Bayesian framework. Several of these will be considered, and the trade-offs they offer will be

[2]However, the results of the work should be transferable to other applications. For example, medical or astronomical images can easily be tackled using the developed methodologies through suitable changes in the observation model, to take account of the different types of observation noise. The reverse may not always hold though, because, as has already been mentioned, some limitations that apply to existing methods may not be suitable to generalise to the wider class of photographic images of natural scenes.

discussed. Along the way, potential hurdles that are faced when restoring real images, such as boundary modelling and computational challenges, will be addressed; ultimately the goal is to derive an algorithm that can be applied to deblur a naturally blurred photographic image.

1.6 Contributions

Several novel contributions are made in the work in this thesis. A new nonstationary image (and blur) model is developed in Chapter 6. Three different BID methods are presented, making use of subsets of this model, in Chapters 7, 9 and 10. A new approach to dealing with boundary conditions is proposed in Chapter 8. A comprehensive review of BID methods is also given, using the Bayesian framework to describe their differences and similarities.

Of the new methods developed, it is felt by the author that the Gibbs sampler method using the full new model, described in Chapter 10 provides a major step forward in restoration quality compared to other published work on BID, especially when dealing with particularly complicated blurs that fail to be estimated correctly with simpler methods. Some examples of the results obtained in Chapter 10 with this method are shown in fig. 1.2.

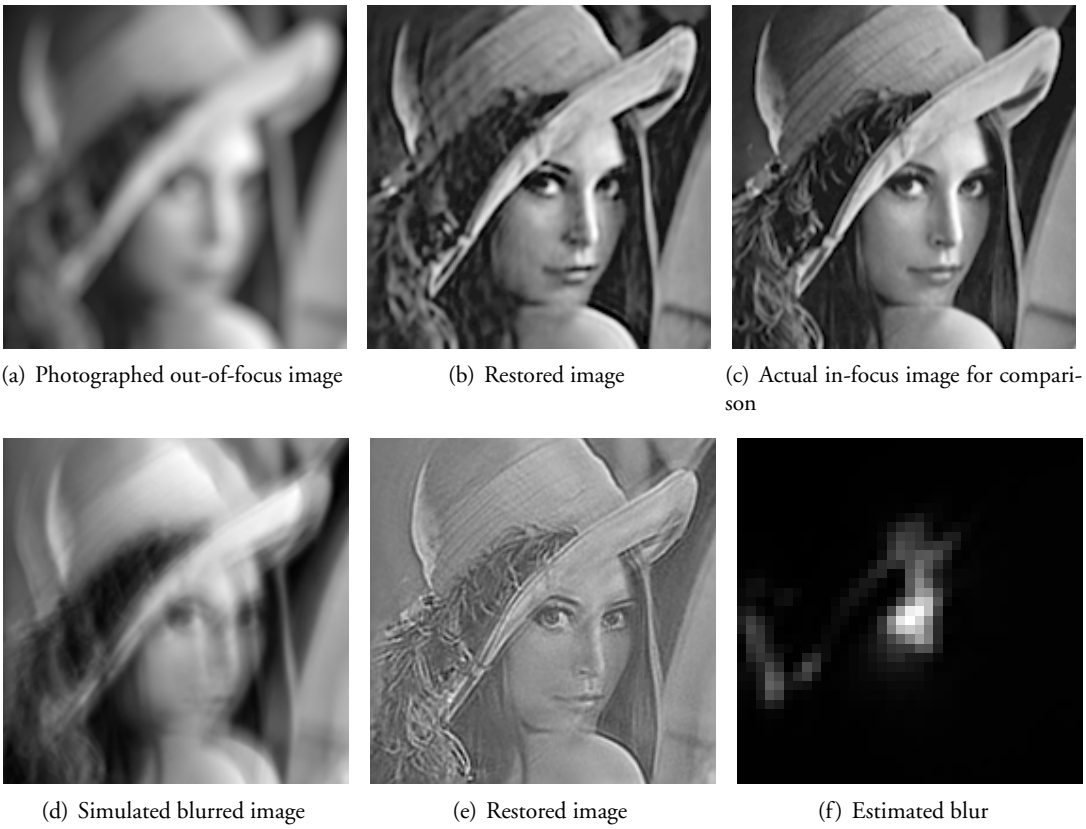
1.7 Outline of Thesis

In this section the contents of each chapter are summarised, and we give suggestions about the recommended reading order. The first part of the thesis represents mainly background material, and an in depth review of both models and methods in the literature. The second part presents the new models that will be used, and several inference methods for solving the BID problem are investigated.

Chapter 2 looks at the commonly used point-spread function (PSF) model of the blurring process, and gives a mathematical formulation of the problem to be solved.

Chapter 3 describes why BID is a difficult task, due to it being an ill-posed problem. Some of the classical image restoration methods and regularisation theory are discussed, including iterative solutions to the deconvolution problem. Some links with other image processing tasks are also mentioned.

Chapter 4 introduces the Bayesian paradigm, upon which this work is built, and surveys the



(a) Photographed out-of-focus image (b) Restored image (c) Actual in-focus image for comparison

(d) Simulated blurred image (e) Restored image (f) Estimated blur

Figure 1.2: *Examples of blind deconvolution using the proposed MCMC method in Chapter 10: (a) – (c) Blind restoration of a real defocused photo; (d) – (f) Blind restoration of an image blurred with a complicated (multi-modal) simulated motion blur.*

prior models for images and blurs. It discusses how inverse problems such as BID naturally fall into the Bayesian framework, and how prior modelling provides a way to incorporate extra available knowledge into the process to get around the ill-posedness of the solution. The hierarchical framework provides a way to extend this to unknowns in the model, and this is discussed, along with other key ideas such as marginalisation and conjugate priors. The various blur and image models that have appeared in the literature which have been used for the solution of BID are then reviewed. Of particular importance are the Markov random field (MRF) image models, as many models can be compared under this representation.

Chapter 5 discusses Bayesian inference techniques and presents a comprehensive review of methods that have appeared in the literature for solving the BID problem. After first considering how to classify these methods in §5.1, the review proceeds by discussing many of the approaches that fall into the Bayesian framework, including some methods which weren't originally presented in this way, but may be reinterpreted as such. This provides a unifying way to draw conclusions about the differences between methods (other methods which cannot be classed in this way are then discussed in Appendix C).

Chapter 6 begins the second part of the thesis, where most of the novel contributions are contained. It takes another look at autoregressive moving average (ARMA) models that have been used in maximum likelihood (ML) blur estimation methods, and uses these concepts to introduce the new nonstationary models that are proposed in this thesis. It shows these models schematically as generative processes, and demonstrates how they can obtain a better fit to the data of a real image than stationary methods. The specification of these probabilistic models based on local AR processes is then given, beginning with a model for zero-mean images in §6.3, then a revised version of this including local means is presented in §6.4. All the additional priors that are used in the hierarchical framework are then specified in the remainder of the chapter, including a model for the blur which takes the same form as the image model.

Chapter 7 presents the first novel method for BID. As an introductory study to the problem and the new models, it considers a simplified version of the full prior model that enables analytic marginalisation of model parameters to be performed. This results in a probabilistic expression for the blur parameters which may be maximised. The marginalisation process reduces the parameter space that has to be searched over. Several experiments using this method are illustrated in §7.2, showing not only the performance of this method with synthetic and

real images, but also how the image model affects the results. Some conclusions are then drawn about why this method is nice in theory but has some difficulties when used with real image degradations.

Chapter 8 takes a slight detour from the path of blind restoration, to deal with Bayesian modelling of the boundary conditions in non-blind image restoration. This is an important topic to consider for any method which aims to deconvolve real blurred images, as errors in the observation model can disastrously affect the results. The methods that are presented are compared experimentally; it is useful to have these methods available to use in the remaining topics of the thesis.

Chapter 9 presents the variational Bayesian (VB) approach to BID, based around approximating the posterior distribution. After reviewing an example based on stationary models for the image and blur, the VB approach using the first new model, which is a simplification of the full model from Chapter 6, using only local modelling of the image variances, is presented in §9.4. While there is some improvement over the stationary model, some problems using this model are shown in §9.4.2. These problems are avoided when also incorporating the local AR model into the estimation procedure in §9.5. Results for blind restoration of Gaussian blurs are shown, and also non-blind restoration in several different conditions helps to validate the quality of the results attainable with this model.

Chapter 10 provides the final novel method, using full simulation of the posterior distribution via the Markov chain Monte Carlo (MCMC) approach of the Gibbs sampler. This is the first time we are aware of this approach having been used to solve the BID problem, partly due to its computational complexity, however in Chapter 10 it is shown that it is able to provide results of impressively high quality. The chapter presents several techniques that had to be combined to ensure reasonably rapid convergence of the simulations. The examples show restorations of several different types of simulated blurs, as well as a real defocused photograph.

There are also several **Appendices**. The first three provide additional background material for the interested reader, while the remaining four provide derivations and additional material related to the novel methods in this thesis. **Appendix A** begins in §A.1 by discussing the notion of the true image, which we wish to restore, and takes a look at models of the photographic imaging process, discussing the various processes which can lead to different types of blur that have to be estimated. In **Appendix B** a more detailed review of the topic of Markov random

fields (MRFs) is given, discussing differences between several variations of the models, followed by a summary of other types of image models. **Appendix C** provides a literature review of non-Bayesian BID methods. **Appendix D** and **Appendix E** contain derivations of results that are used for the MMAP method and boundary models in Chapter 7 and Chapter 8 respectively, while **Appendix F** and **Appendix G** describe derivations that are used for the distribution approximations in Chapter 9. Finally Appendix H lists the contents of the included DVD, which contains video clips showing the progress of the restorations from Chapters 9 and 10.

The reader who is familiar with the BID topic, image modelling and Bayesian methods, may skip directly to Part II. Otherwise it is suggested to at least read §2.1 to be familiar with the problem, and Chapter 4 and §5.2 on prerequisites for the use of Bayesian methods. Chapter 3 may also be useful for the reader without knowledge of inverse problems or image restoration. In Part II, it is advisable to begin with Chapter 6 which describes the models used in later chapters. After this, the other topics are more or less independent and can be read in any order.

Chapter 2
Problem formulation

Before we begin trying to solve the problem of BID, we should examine what types of degradations we are likely to encounter in practice. In other words, we should model the photographic observation process. A detailed analysis is presented in Appendix A, however we begin here by assuming that we can approximate all these effects using a linear system.

2.1 Basic Mathematical formulation of BID

The model that is most commonly used to represent the observation process in the BID problem is a discrete linear spatially-varying (LSV) or LSI convolution model. The images are defined here as discrete 2D signals. We begin by denoting by $f(\mathbf{s})$ the *true* or *original* image (the terms will be used interchangeably, see §A.1 for further discussion), which is the main quantity of interest. The observed image will be denoted $g(\mathbf{x})$, and the kernel function describing the blur as $h(\mathbf{x}, \mathbf{s})$. The indices $\mathbf{x} = (x_1, x_2)$ and $\mathbf{s} = (s_1, s_2)$ represent pixel locations on the infinite regular 2D lattices, or canvases, $\Omega_s \equiv \mathbb{Z}^2$ and $\Omega_x \equiv \mathbb{Z}^2$. It is useful to think of $\mathbf{s}$ and $\mathbf{x}$ as lying in two distinct co-ordinate spaces, in the input and output planes to the imaging system, respectively (fig. 2.1).

The discrete image f is defined at pixel locations $\mathbf{s}$ contained within a support region $\mathcal{S}_f \subset \Omega_s$, and similarly g has pixels defined within valid region of support $\mathcal{S}_g \subset \Omega_x$; outwith these regions the signals are assumed to be equal to zero. This notation will be elaborated upon when discussing MRF models in Appendix B and boundary conditions in §8.2. The support regions will be assumed to be rectangular, with $\mathcal{S}_f = \{1 \cdots m_f\} \times \{1 \cdots n_f\}$ and $\mathcal{S}_g = \{1 \cdots m_g\} \times \{1 \cdots n_g\}$.

The LSV observation model is now stated mathematically as:

$$g(\mathbf{x}) = [\, h(\mathbf{x}, \mathbf{s}) \star f(\mathbf{s}) \,](\mathbf{x}) + w(\mathbf{x}) \tag{2.1}$$

$$= \sum_{\substack{\mathbf{s} \in \mathcal{S}_f, \\ (\mathbf{x}, \mathbf{s}) \in \mathcal{S}_H}} h(\mathbf{x}, \mathbf{s}) f(\mathbf{s}) + w(\mathbf{x}), \qquad \mathbf{x} \in \mathcal{S}_g. \tag{2.2}$$

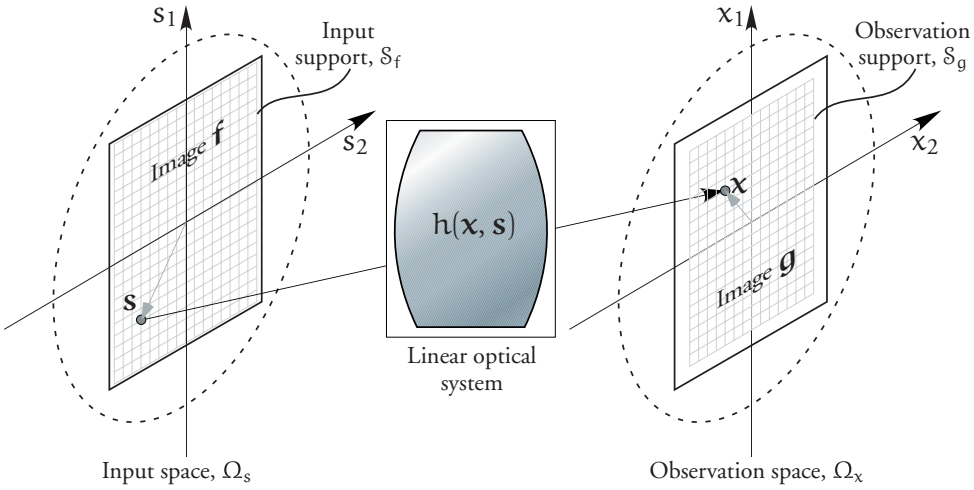


Figure 2.1: *Kernel representation of the LSV observation model. The observed response at point* $\mathbf{x}$ *to a point source at* $\mathbf{s}$ *in the input image is weighted according to the kernel* $h(\mathbf{x}, \mathbf{s})$.

The symbol $\star$ denotes (generalised) convolution. The term $w(\mathbf{x})$ represents additive noise that corrupts the observations. The kernel, or spatially-varying PSF, $h(\mathbf{x}, \mathbf{s})$, parameterises the imaging system, relating points in the input space to the observation space. It is the 2D equivalent of an impulse response, describing the response of the system at position $\mathbf{x}$ in the output space to an impulse or point source at location $\mathbf{s}$ in the input space.

Thus $h(\mathbf{x}, \mathbf{s})$ may be defined in the domain $\Omega_x \times \Omega_s$, for all combinations of $\mathbf{x}$ and $\mathbf{s}$ within the support regions $\mathcal{S}_g$ and $\mathcal{S}_f$; though in practice it may be zero for many combinations where $r = |\mathbf{s} - \mathbf{x}|$ is sufficiently large, indicating that $f(\mathbf{s})$ does not influence $g(\mathbf{x})$ at this location. Then the summation may be reduced to exclude these points and include only those pairs of points $(\mathbf{x}, \mathbf{s})$ within the support $\mathcal{S}_H$ of the kernel $h(\mathbf{x}, \mathbf{s})$.

2.1.1 Spatially Invariant model

In practice, the general spatially varying (SV) form is often too complicated to handle computationally, and also results in an overwhelming number of parameters, so the simplification of a spatially invariant (SI) blur is often used (methods for addressing the SV problem are considered in §C.8). This assumes that the response at $\mathbf{x}$ to a point at $\mathbf{s}$ only depends on their relative displacement $\mathbf{r} = \mathbf{x} - \mathbf{s}$, or equivalently, a PSF $h(\mathbf{r})$ may be introduced. This PSF is a function defined in a domain $\Omega_r \equiv \mathbb{Z}^2$, with a support $\mathcal{S}_h = \{1 \cdots m_h\} \times \{1 \cdots n_h\}$, and is

independent of position in the image. The LSI model using this notation is then written as

$$g(\mathbf{x}) = [h \star f](\mathbf{x}) + w(\mathbf{x}) \tag{2.3}$$

$$= \sum_{\substack{\mathbf{s}\in\mathcal{S}_f,\\ \mathbf{x}-\mathbf{s}\in\mathcal{S}_h}} h(\mathbf{x}-\mathbf{s})f(\mathbf{s}) + w(\mathbf{x}), \qquad \mathbf{x}\in\mathcal{S}_g \tag{2.4}$$

$$= \sum_{\substack{\mathbf{r}\in\mathcal{S}_h,\\ \mathbf{x}-\mathbf{r}\in\mathcal{S}_f}} f(\mathbf{x}-\mathbf{r})h(\mathbf{r}) + w(\mathbf{x}), \qquad \mathbf{x}\in\mathcal{S}_g. \tag{2.5}$$

The simplified LSI model is shown as a block diagram in Fig. 2.2.

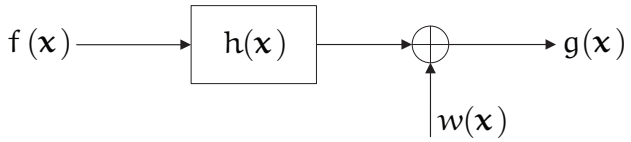


Figure 2.2: *Basic LSI observation model*

2.1.2 Matrix-vector formulation

More commonly, the images and PSF are represented by rearranging their pixels into 1D vectors. This may be via lexicographic ordering (see fig. 2.3) or some other suitable ordering of the pixel indices $\mathbf{x}$, $\mathbf{s}$, and $\mathbf{r}$ (e.g. in §6.3.4 we use a blockwise scan). Lexicographic ordering will be denoted as, for instance, $\mathbf{f} = \mathrm{vec}(f(\mathbf{s}))$, and the reverse process as $f(\mathbf{x}) = \mathrm{unvec}(\mathbf{f})$. (2.4) and (2.5) may then be rewritten in a compact matrix-vector form:

$$\mathbf{g} = \mathbf{H}\mathbf{f} + \boldsymbol{w}, \qquad \text{or equivalently,} \tag{2.6}$$

$$\mathbf{g} = \mathbf{F}\mathbf{h} + \boldsymbol{w}. \tag{2.7}$$

With the spatially-invariant degradation model that is assumed above, and lexicographic ordering of the images, the matrices $\mathbf{F}$ and $\mathbf{H}$ acquire a special structured block form, termed block Toeplitz with Toeplitz blocks (BTTB) (also known as Doubly Block Toeplitz in [96]). This form has constant block entries on each block diagonal and constant diagonals within each block, such that $\mathbf{H}_{\mathbf{x},\mathbf{s}} = h(\mathbf{x}-\mathbf{s})$ and $\mathbf{F}_{\mathbf{x},\mathbf{r}} = f(\mathbf{x}-\mathbf{r})$.[1] Note that for unequal support sizes $\mathcal{S}_f$ and $\mathcal{S}_g$ of the images, $\mathbf{F}$ and $\mathbf{H}$ will be non-square. We will use the notation $L_f = |\mathcal{S}_f| = m_f n_f$,

[1] The following subscript notation for indexing entries in a matrix representing a 2D convolution operation between lexicographically ordered vectors will be used for simplicity: $\mathbf{H}_{\mathbf{x},\mathbf{s}} \triangleq \mathbf{H}_{(n_g-1)x_1+x_2,(n_f-1)s_1+s_2}$, i.e. the entry of the matrix at row $(n_g-1)x_1+x_2$ and column $(n_f-1)s_1+s_2$.

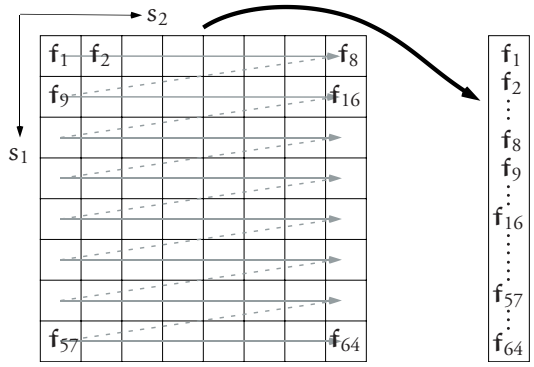


Figure 2.3: *Lexicographic ordering, or raster scanning of the image. The* 8×8 *image* $f(\mathbf{s})$ *on the left is reordered to give the vector* $\mathbf{f}$ *on the right.*

$L_g = |\mathcal{S}_g| = m_g n_g$, and $L_h = |\mathcal{S}_h| = m_h n_h$ to denote the length of the relevant vectors.

The spatially varying case with the kernel representation described in (2.2) can also be rewritten in matrix form as (2.6). Again we let $\mathbf{f} = \mathrm{vec}(f(\mathbf{s}))$ and $\mathbf{g} = \mathrm{vec}(g(\mathbf{x}))$, but now $\mathbf{H}$ is no longer BTTB. Instead it is defined with entries $\mathbf{H}_{\mathbf{x},\mathbf{s}} = h(\mathbf{x}, \mathbf{s})$.

2.1.3 Interpretation and visualisation of 2D convolution

In order to visualise what is happening in the convolution operation more clearly (especially in the SV case), two alternative representations of the model will be described. These are essentially different manners in which to interpret (2.6). We will introduce the *spatially varying PSF*, $h_{\mathbf{s}}(\mathbf{r}_s)$, and the *spatially varying PIF*, $h_{\mathbf{x}}(\mathbf{r}_x)$. The vector $\mathbf{r}_s(\mathbf{x}) \triangleq \mathbf{x} - \mathbf{s}$ defines a co-ordinate system in Ω_x centred at the pixel $\mathbf{s}$ as a function of $\mathbf{x}$, and likewise $\mathbf{r}_x(\mathbf{s}) \triangleq \mathbf{s} - \mathbf{x} = -\mathbf{r}_s(\mathbf{x})$ one in Ω_s centred on a particular pixel $\mathbf{x}$ as a function of $\mathbf{s}$. In these co-ordinate systems, the functions $h_{\mathbf{s}}(\mathbf{r}_s)$ and $h_{\mathbf{x}}(\mathbf{r}_x)$ are defined within supports of $\mathcal{S}_h(\mathbf{s})$ and $\mathcal{S}_h(\mathbf{x})$ respectively.

With these definitions, the LSV model can also be written as:

$$g(\mathbf{x}) = \sum_{\substack{\mathbf{s}\in\mathcal{S}_f,\\ \mathbf{r}_x(\mathbf{s})=(\mathbf{s}-\mathbf{x})\in\mathcal{S}_h(\mathbf{x})}} h_{\mathbf{x}}(\mathbf{x}-\mathbf{s}) f(\mathbf{s}) + w(\mathbf{x}), \qquad \mathbf{x} \in \mathcal{S}_g. \tag{2.8a}$$

$$= \sum_{\substack{\mathbf{s}\in\mathcal{S}_f,\\ \mathbf{r}_s(\mathbf{x})=(\mathbf{x}-\mathbf{s})\in\mathcal{S}_h(\mathbf{s})}} h_{\mathbf{s}}(\mathbf{x}-\mathbf{s}) f(\mathbf{s}) + w(\mathbf{x}), \qquad \mathbf{x} \in \mathcal{S}_g. \tag{2.8b}$$

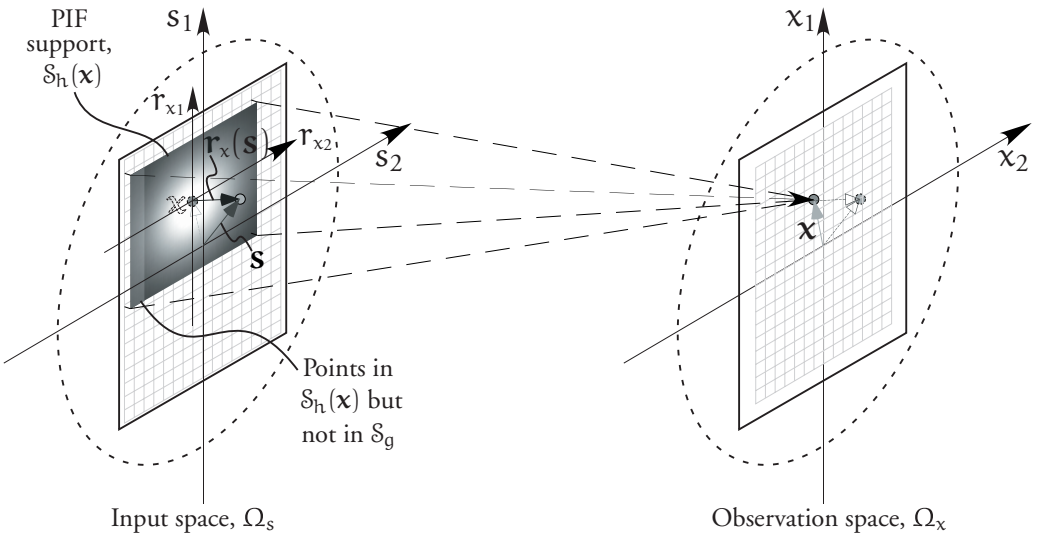


Figure 2.4: *The point-influence function (PIF) representation of the convolution process. The points within the region* $\mathcal{S}_h(\mathbf{x})$ *on the left are weighted according to the function* $h_x(-\mathbf{r}_x)$*, and summed to give the response at* $\mathbf{x}$ *in the observed image.*

Observe the similarity to the LSI form in (2.6), which is recovered if the dependencies on $\mathbf{x}$ and $\mathbf{s}$ of h and $\mathcal{S}_h$ are dropped. In general, $h_{\mathbf{x}}(\mathbf{r}_x) \neq h_{\mathbf{s}}(\mathbf{r}_s)$, but in the SI case we have that $h(\mathbf{r}) = h_{\mathbf{s}}(\mathbf{r}_s) = h_{\mathbf{x}}(-\mathbf{r}_x)$. In this case also $\mathcal{S}_h$ is stationary, and we assume a rectangular support of size $m_h \times n_h$.

Considering the first case, (2.8a) , the convolution process can be thought of visually as each point in the image $g(\mathbf{x})$ being influenced by a sum of its neighbours in $f(\mathbf{s})$, weighted by the PIF $h_{\mathbf{x}}(-\mathbf{r}_x)$ at the appropriate spatial shift $\mathbf{r}_x$, as shown in fig. 2.4. In terms of the vector-matrix formulation, this corresponds to each element of $\mathbf{g}$ being the projection of $\mathbf{f}$ onto the corresponding row of of $\mathbf{H}$. That is to say, row $\mathbf{x}$ of $\mathbf{H}$ is given by $\text{vec}(h_{\mathbf{x}}(-\mathbf{r}_x))$. Observe that the PIF is mirrored before translating to the co-ordinate frame $r_x(\mathbf{s})$.

Alternatively, in (2.8b), the whole image $g(\mathbf{x})$ can be seen as a summation of PSFs $h_{\mathbf{s}}(\mathbf{r}_s)$ centred at and weighted by each pixel $\mathbf{s}$ in the image $f(\mathbf{s})$, as shown in fig. 2.5. This is represented in matrix-vector form by summing the columns of $\mathbf{H}$ each multiplied by the corresponding entries of $\mathbf{f}$, to give $\mathbf{g}$. Column $\mathbf{s}$ of $\mathbf{H}$ is given by $\text{vec}(h_{\mathbf{s}}(\mathbf{r}_s))$.

The coordinate $\mathbf{r}$ in each case defines the distance away from the diagonal or block-diagonal

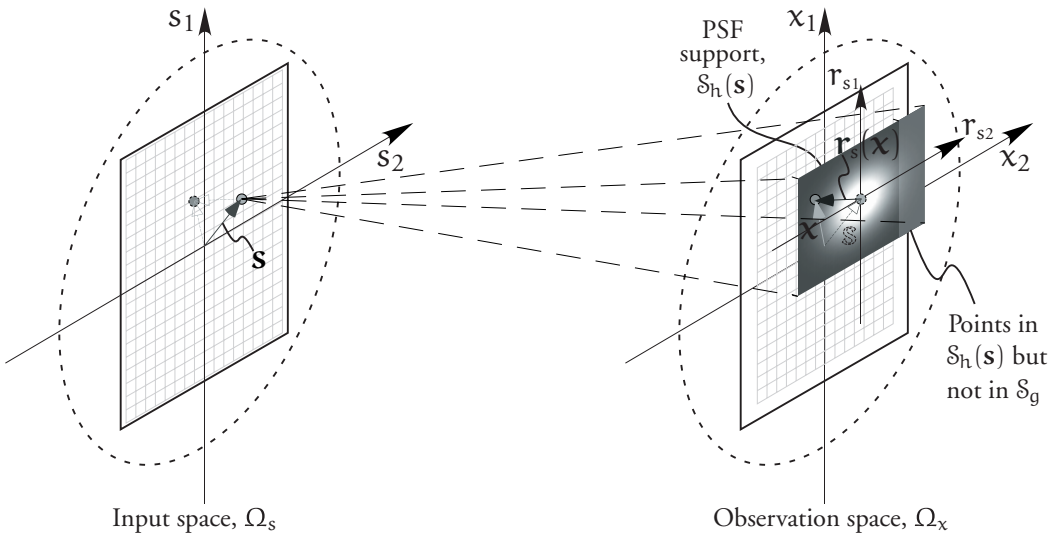


Figure 2.5: *The point-spread function (PSF) representation of the convolution process. A pixel at point* $\mathbf{s}$ *in the true image spreads out with a response* $h_s(\mathbf{r}_s)$*, affecting those points in* $\mathcal{S}_h(\mathbf{s})$ *lying within the observation region* $\mathcal{S}_g$*. All such responses may be summed across the points* $\mathbf{s}$ *in* $\mathcal{S}_f$ *to give the observed response to an arbitrary image.*

of the matrix.

The two interpretations just described correspond to equations (2.4) and (2.6) in the stationary case. Similar visualisations may be obtained for the case where the role of image and blur are swapped, as in (2.5) or (2.7). In the stationary case, the PIF type representation again holds from projecting $\mathbf{h}$ onto the rows of $\mathbf{F}$; and considering $\mathbf{g}$ as the sum of the columns of $\mathbf{F}$ multiplied by the corresponding entries of $\mathbf{h}$ reveals a PSF type representation. The difference now is that rather than moving the PSF around and projecting onto the image, the PSF coordinate system $\mathbf{r}$ is fixed in space, and copies of the image f are made at $(\mathbf{x} - \mathbf{r})$, which are weighted by the PSF coefficients, and these summed to give the observation as before.

In the non-stationary case, the vector $\mathbf{h}$ contains all the responses $\text{vec}(h_s(\mathbf{r}_s))$ for $\mathbf{s} \in \mathcal{S}_f$, stacked. Matrix $\mathbf{F}$ becomes an expanded block diagonal version of the stationary version, and only the PIF representation really makes sense. With the PSF representation, each column of the $\mathbf{F}$ matrix would only contain a single pixel, thus rendering the interpretation much the same as the kernel representation. A block-stationary case is still useful to consider, and this will be discussed again in the context of image modelling in Chapter 6.

2.1.4 Problem specification

Equations (2.4), (2.5), (2.6), or (2.7) help to mathematically specify the problem at hand. The classical image restoration task is to estimate the image $\mathbf{f}$ from the observation $\mathbf{g}$, given knowledge of $\mathbf{h}$. The BID problem that will be addressed in this thesis considers the more complicated problem of estimating $\mathbf{f}$ from $\mathbf{g}$ when $\mathbf{h}$ is also unknown.

Classically, estimators are used that provide point estimates $\hat{\mathbf{f}}$ and $\hat{\mathbf{h}}$, assuming there is only one true value of the unknowns. However in the Bayesian paradigm we wish to infer the range of *likely* values these variables could take, thus the quantities of interest are their probability density functions (PDFs), conditional on the observation, $p(\mathbf{f} \mid \mathbf{g})$ or $p(\mathbf{h} \mid \mathbf{g})$. In practice we may choose to take the mean or mode of these distributions as a point estimate representing the solution, and then specify the uncertainty in the form of the distributions' variances.

2.2 Conclusions

In this chapter, the linear systems representation of the observation process was discussed, leading to the standard PSF model of spatially invariant blurring, which we will use in representing the BID problem we wish to solve. The SV image formation case was also discussed, where there are two equivalent representations, the PSF model and a model we have termed the PIF model; essentially these are different ways of writing the convolution operation, or thinking about the linear system. In the next chapter we look at how the convolution process and linear algebra can lead to the idea of ill-posed, or ill-conditioned problems.

Chapter 3
Ill-posed problems and Classical restoration

3.1 Ill-posed Inverse problems

As an introduction to the challenges that are faced in blind restoration, the topic of ill-posed inverse problems and regularisation will be considered; in particular, their application to the field of classical image restoration, where the PSF is assumed known, and the task is to recover the true image from a blurred and noisy observation. These problems will be considered using the classical approach of regularisation theory, before we later examine this as a special case of the Bayesian framework.

These topics provide the foundations for many blind deconvolution techniques. This material is covered in a basic form in many image processing textbooks [79, 96, 179], and in some more depth in [5] and [18]. A very good overview of the main techniques used to solve ill-conditioned problems is given in [101]. Reviews of the image restoration field have been given, for example, in [12, 120, 199] and even earlier work in [202].

Mathematically, image restoration requires solving an ill-posed problem, which in the continuous case may be stated in the form of a Fredholm integral of the first kind with the addition of a noise term:

$$g(x,y) = \int_{-\infty}^{\infty}\int h(x,y;x',y')f(x',y')\,dx'\,dy' + w(x,y). \tag{3.1}$$

Solution of (3.1) for f constitutes an inverse problem, because we are trying to invert the naturally occurring process. A detailed account of ill-posedness in inverse problems is given in [18]. In the continuous case, the integral of (3.1) defines a mapping $\mathcal{H}$ from an object space $\mathcal{F}$ to an image space $\mathcal{G}$ (a superset of the range space of blurred images). The direct, or forward problem, from $\mathcal{F}$ to $\mathcal{G}$ usually represents a physically realisable process, whereas the inverse problem from $\mathcal{G}$ to $\mathcal{F}$ represents that that we wish to solve.

Inverse problems are *well-posed* (a concept due to Hadamard in the early 1900's) if the solution depends continuously on the data, in a smooth manner. This implies small measurement errors do not significantly change the solution of the inverse problem. On the contrary, the mapping defined by an *ill-posed* problem may be non-unique or a discontinuous function of the data, representing loss of information, giving a difficulty for the inverse problem. Two wildly different objects may have the same image (possible for example in a perspective projection), or there may be a large region in the object space that maps to a very small region of image space (often the case with noise processes), thus an arbitrarily small perturbation of the measurement (i.e. noise) can yield an arbitrarily large error in the restoration. The exact solution may in fact be non-unique if the mapping is degenerate, or even not exist if the noisy observation does not lie in the range space of $\mathcal{H}$.

Of course in practise a discrete model is used for the sampled digital image on a regular 2D lattice, such that the integrals are replaced by summations, giving (2.2). In this discrete case, ill-posedness becomes *ill-conditioning* of the matrix equation (2.6). It is possible that there may now be a single exact solution to the problem, yet due to the ill-conditioning and noise it may not be near the solution from the noiseless observation: it may be mathematically correct yet physically meaningless.

The degree of ill-conditioning is represented by the condition number, the ratio of the largest to smallest singular values. We observe that for small blurs (when the degree of discretisation is coarse) the condition number is small; however it tends to infinity as the blur size or fineness of discretisation increases, due to approaching the ill-posed continuous problem, and due to machine imprecision (the small singular values are rounded to zero, and $\mathbf{H}$ becomes singular).

In practice, the problem is often worsened by the fact that the small singular values usually represent high spatial frequencies in the image, and the corrupting noise typically dominates over the image at high spatial frequencies. Therefore it is likely this high-frequency noise pattern will overwhelm the inverse solution. Many blurs also contain zeros in their spectra; when these are inverted the solution becomes infinite at these frequencies which is obviously physically unrealistic.

3.1.1 Inverse solution

The simplest solution is the inverse filter. In the case of a non-square matrix,[1] where there may not exist an exact solution to the matrix equation (inconsistent data, singular or underdetermined system), the least squares (LS), also known as generalised- or pseudo-inverse formulation may be used. This seeks to minimise the squared residual error norm, $\|\mathbf{g} - \mathbf{H}\hat{\mathbf{f}}\|^2$ between the observation, and its projection, $\hat{\mathbf{g}} = \mathbf{H}\hat{\mathbf{f}}$, into the space of blurred images — the column space of blur matrix. The error lies in the left-nullspace of the blur matrix, and represents the noise component incompatible with the assumed blur model (assuming the space of blurred images does not span the space of all images in $\mathbb{R}^{m\times n}$, i.e. $\mathbf{H}$ is not full rank). Minimisation may be accomplished by differentiating the error with respect to (w.r.t) $\hat{\mathbf{f}}$, yielding

$$\mathbf{0} = \mathbf{H}^{\mathsf{T}}\mathbf{g} - \mathbf{H}^{\mathsf{T}}\mathbf{H}\hat{\mathbf{f}} \tag{3.2}$$

$$\Rightarrow \hat{\mathbf{f}} = \mathbf{H}^{+}\mathbf{g} = (\mathbf{H}^{\mathsf{T}}\mathbf{H})^{-1}\mathbf{H}^{\mathsf{T}}\mathbf{g} \tag{3.3}$$

(This may also be evaluated in the discrete frequency domain using the methods of §3.2.1). Unfortunately additive noise, which is not explicitly modelled by this solution, tends to be amplified by the psuedoinverse as $\mathbf{H}^{+}\boldsymbol{w}$. Due to the ill-conditioning already mentioned, this will tend to give a large deviation to the solution (amplified noise) in the row space of $\mathbf{H}$. If $\mathbf{H}$ is singular then there exists a nullspace containing images that are blurred to give a zero image. There will be a range of solutions containing the "desired" solution plus an additional random noise component from the nullspace of $\mathbf{H}$ (it is possible too that the original image contained a nullspace component, which will be unrecoverable). Very small singular values represent basis images that blur to a "near zero" image, i.e. almost in the nullspace, and in the inverse these represent amplification of perturbations to large noise in the row space. These concepts are displayed pictorially in fig. 3.1, representing the two pairs of orthogonal vector subspaces and the true and observed image spaces.

Thus in order to tackle the ill-conditioned problem we require methods to obtain a physically meaningful approximate solution given the data, with some additional constraints that represent our beliefs or *a priori* knowledge about what such a physically meaningful solution should be. This consists of constraining the solution to a smaller set, typically considering the energy, boundedness, or smoothness of the solution, but may include a more complicated model of

[1] Note the matrix $\mathbf{H}$ will always be square if the input and output images are the same size, in which case the LS formulation reduces to the inverse filter; if we attempt to estimate a larger image size than the observation an underdetermined system results, if we only wish to estimate the central part we have an overdetermined system.

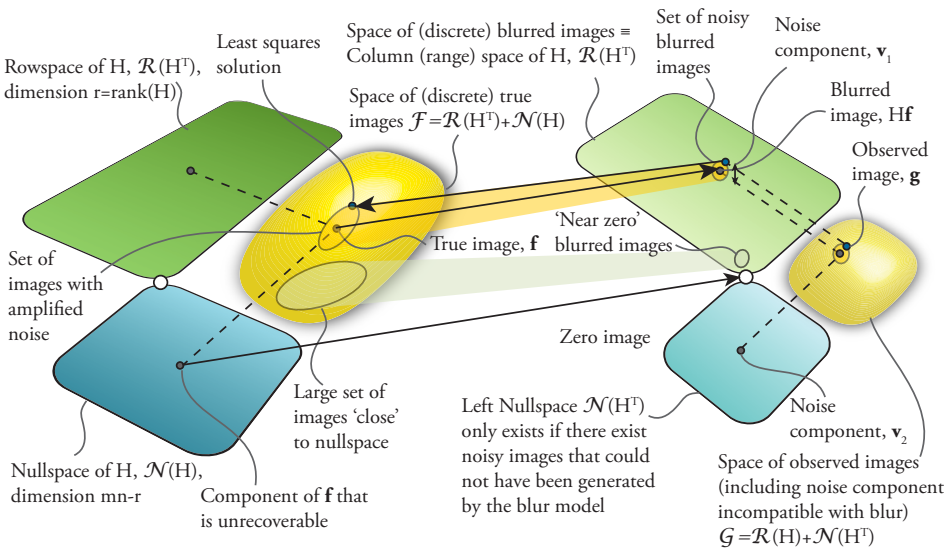


Figure 3.1: *Vector space representation of Ill-conditioned problem*

the data we are looking for. It may also be seen as instead solving an associated well-posed problem. In general these techniques are called *regularisation* methods; their initial application is generally attributed to Tikhonov [211]. Regularisation may also be viewed as an example of Bayesian techniques, where the constraints are provided in stochastic form via the prior PDF.

3.1.2 Regularised methods

Some basic ad-hoc techniques to overcome the problems with inverse filtering are mentioned in [202], however the most common early regularisation methods include the constrained least squares (CLS) filter [95] and the Wiener filter.

3.1.2.1 Constrained Least Squares

The CLS filter modifies the original LS formulation by adding a smoothness constraint, or *stabilising functional*, such that the energy of $\|\mathbf{C}\hat{\mathbf{f}}\|^2$ is to be minimised in addition to the residual from before. $\mathbf{C}$ in general represents a high-pass action on the image, penalising non-smooth images (typically the discrete approximation to the 2D Laplacian is used [79, 95]). This may be achieved via the method of Lagrange multipliers, which converts the constrained

problem into an unconstrained one:

$$\hat{\mathbf{f}} = \arg\min_{\hat{\mathbf{f}}} \left(\|\mathbf{g} - \mathbf{H}\hat{\mathbf{f}}\|^2 + \alpha \|\mathbf{C}\hat{\mathbf{f}}\|^2 \right) \tag{3.4}$$

$$\Rightarrow \hat{\mathbf{f}} = (\mathbf{H}^{\mathsf{T}}\mathbf{H} + \alpha \mathbf{C}^{\mathsf{T}}\mathbf{C})^{-1}\mathbf{H}^{\mathsf{T}}\mathbf{g}, \tag{3.5}$$

where the solution is again found by differentiation w.r.t $\hat{\mathbf{f}}$.

The regularisation parameter, α, trades off the goodness of fit of the original versus the smoothness constraint. Some methods for its choice are compared in [70]. It may be chosen in an interactive fashion, by assuming a value for the residual error based on estimated image / noise statistics; the Miller method (assuming bounds on the magnitude of the residual and penalty, which is connected to the set-theoretic approach, giving a set of admissible solutions as the intersection of two ellipsoids in energy space [109]); or automatically by methods such as generalised cross-validation (GCV) [186] (also known as "leave one out" — the data is divided into a estimation and validation sets, and different parameters tried to minimise the error in the validation set using the solution obtained from the estimation set); or the L-Curve method [88] (which plots the residual versus the smoothness term, the resulting corner of the L shaped graph dictates the optimum parameter). As will be seen in later chapters, the hierarchical Bayesian approach provides a natural solution to estimation of regularisation parameters as part of the inference procedure.

The mathematical properties of the degree of deblurring possible in an ill-conditioned problem have been analysed in the form of the discrete Picard condition (DPC) [87], which considers the spectrum of singular values of the blur; this is also analysed for the related problem of superresolution (SR) in [222].

3.1.2.2 Wiener filter

The Wiener filter is somewhat similar to the CLS method but uses stochastic rather than deterministic regularisation principles. Here we seek to minimise the mean squared error expectation:

$$\hat{\mathbf{f}} = \arg\min_{\hat{\mathbf{f}}} \mathrm{E}\left[\|\mathbf{f} - \hat{\mathbf{f}}\|^2 \right] \tag{3.6}$$

$$\Rightarrow \hat{\mathbf{f}} = (\mathbf{H}^{\mathsf{T}}\mathbf{H} + \mathbf{R}_{\mathrm{ff}}^{-1}\mathbf{R}_{ww})^{-1}\mathbf{H}^{\mathsf{T}}\mathbf{g}, \tag{3.7}$$

where the solution is found by differentiation, and we require knowledge of the image and noise second order statistics, $\mathbf{R}_{\mathbf{ff}} = \mathsf{E}[\mathbf{f}\mathbf{f}^{\mathsf{T}}]$ and $\mathbf{R}_{ww} = \mathsf{E}[\boldsymbol{w}\boldsymbol{w}^{\mathsf{T}}]$. Unfortunately this is difficult in practice and they must be estimated from the image itself, so the CLS solution is more commonly used (their performance is fairly similar). Once more, implementation in the frequency domain (using the power spectra in place of the covariances) is common.

Note that these filters may also be derived from a Bayesian perspective. For example, with the Wiener filter, the likelihood comes from the PDF of the noisy observations, which are assumed to be Gaussian with covariance $\mathbf{R}_{ww}$ and mean $\mathbf{H}\mathbf{f}$, and the prior term encapsulates the PDF of the true image, which is assumed to be a zero-mean Gaussian with covariance $\mathbf{R}_{\mathbf{ff}}$. By finding the mean of the posterior PDF for $\mathbf{f}$, we arrive at the Wiener filter (see §4.2).

Both of these methods work using mean squared error (MSE) as an optimality measure; the Wiener filter is the optimal filter for linear restoration. However the linear filters tend to exhibit artefacts such as ringing at edges (Gibbs oscillations) due to partial deblurring, incompatible with our expectations of reality. It has been acknowledged, even in early work [202], that MSE is not the best criterion for image fidelity. A metric that appeals to the qualities of the human visual system (HVS) would be more appropriate (see §3.2.5). Furthermore, better performance may be gained through use of non-linear processing, and incorporating a more accurate model of the underlying image and degradation process.

The numerical implementation of inverse methods is also very important, and has been the subject of much research. Due to the size of the matrices needed for inversion, direct solution of (3.5) and (3.7) is usually prohibitive; for example, 256×256 pixel images $\mathbf{f}$ and $\mathbf{g}$ result in a matrix H which requires 32GB of memory for double precision. If the support of the PSF is small, it is possible to store the matrix using a sparse representation and use sparse linear algebra solvers to find the solution. However there are often faster methods using approximations; some common techniques will be described in the next two sections.

3.2 Classical Image restoration methods

3.2.1 Circulant approximations

One common approximation used to dramatically reduce the storage and processing requirements is to assume a periodic image boundary, or a toroidal lattice, giving a block circu-

lant with circulant blocks (BCCB) structured $\mathbf{H}$ representing circular convolution. This enables important speed gains to be realised by means of the 2D FFT. It can be shown [5, 21, 79, 94, 178, 179] by an eigenvector analysis that this transform is exactly that which diagonalises the linear system of $\mathbf{H}$; the eigenvalues of the BCCB $\mathbf{H}$ are the 2D DFT coefficients of the shifted PSF image, $\mathfrak{h}(\mathbf{r})$, and the eigenvectors the 2D DFT basis functions. This enables a massive speedup and reduction in memory requirements to be realised (to the extent that early computer processing was impossible without this technique).

Further justification for the approximation is given by the equivalent (in the sense of the Euclidean norm) asymptotic behaviour of Circulant and Toeplitz matrices, due to theorems by Szegö considering Eigen- or Singular-Value distributions, which have been discussed in [81, 82]; recently similar theory for BCCB and BTTB matrices has been exposed in [30], along with discussions of relevance to preconditioning methods (see §3.2.4.2).

Careful consideration of the effects of lattice boundary conditions with the finite model are very important however. In practice we do not have a circular boundary and there is in fact a missing data problem at the edges where energy has been lost outside the recorded image (and we have some "bleed-in" from the unknown outside). [79] suggests that the BTTB form of (2.6) (for an $\mathfrak{m}\times\mathfrak{n}$ image) of an $\mathfrak{mn}\times\mathfrak{mn}$ matrix, $\mathbf{H}$, representing a blur, $\mathfrak{h}$ of size $\mathfrak{m}_\mathfrak{h}\times\mathfrak{n}_\mathfrak{h}$ may be "embedded" in a BCCB one by first zero-padding the image data to size $(\mathfrak{m}+\mathfrak{m}_\mathfrak{h})\times(\mathfrak{n}+\mathfrak{n}_\mathfrak{h})$ and forming the new $\mathbf{H}$ as the BCCB matrix defined by $\mathfrak{h}$. Then the result of the convolution using the discrete Fourier transform (DFT) is identical to the exact result, ignoring the zero padded regions. This procedure holds for the forwards process, however it does not carry over to the deconvolution process in the discrete frequency domain. This is partly due to the infinite support of the inverse filter and partly due to the data truncation that happened in the forwards process. An analysis of these problems is given in [86].

The assumed periodic image plane introduces artificial high frequencies due to the unnatural step transition at the edge of the blurred image. These will then be amplified by the restoration filter and manifest as extreme ringing artefacts in the restored image. Ringing artefacts also result at natural edges in the image due to partial deblurring (see fig. 3.3(a); some high frequencies have not been fully restored), however the boundary edge effects will be far worse since the restoration filter is attempting to deblur an already sharp edge. Some methods to deal with this have been proposed: the Dirichlet boundary condition (BC) [31] (assume zeros outside data), use of a reflective tiling (the Neumann BC [48, 163], reportedly better than the

Dirichlet BC), to interpolate the image borders [120], or the method used in [23] of weighting the data at the boundaries according to the PSF. Boundary conditions are also discussed in [5].

A more careful examination of these boundary conditions and how to effectively deal with them under the Bayesian framework will be presented in Chapter 8.

3.2.2 Iterative methods

Though the direct evaluation of (3.5) or (3.7) is prohibitive due to the inversions involved, and while FFT methods may ease computational burden, with the provisos mentioned, there is another possibility. (3.4) may instead be solved by application of iterative methods [21, 109] such as steepest descent (SD) or conjugate gradients (CG). This provides a number of advantages: the possibility of non-linear or spatially adaptive restoration (avoiding boundary condition problems), coping with SV blurs and image statistics; matrix inversions are usually not required; parameters may be adjusted adaptively or iterations halted early to provide an additional degree of regularisation.

Applying SD (or the method of successive approximations) to the original problem, $\mathbf{g} = \mathbf{H}\mathbf{f}$, results in the basic *van Citert* iteration, $\hat{\mathbf{f}}_{i+1} = \hat{\mathbf{f}}_i + \beta(\mathbf{g} - \mathbf{H}\hat{\mathbf{f}}_i)$. It can be shown by eigenvalue analysis that under certain restrictive conditions (e.g. for blurs with positive real parts in their spectra, and with the correct choice of the convergence parameter β) this will converge to the inverse filter (but with the advantage of early termination of the iterations which reduces noise amplification). If instead the LS formulation $\mathbf{H}^T\mathbf{g} = \mathbf{H}^T\mathbf{H}\mathbf{f}$ is minimised, the re-blurred or *Landweber* method, $\hat{\mathbf{f}}_{i+1} = \hat{\mathbf{f}}_i + \beta\mathbf{H}^T(\mathbf{g} - \mathbf{H}\hat{\mathbf{f}}_i)$, is obtained, which is slightly less restrictive, converging (correctly for blurs with no zero eigenvalues) to the pseudoinverse solution. The most general form is the *Tikhonov-Miller* solution which is essentially the iterative version of the CLS filter; again by applying SD to (3.4), the result is:

$$\hat{\mathbf{f}}_{i+1} = (\mathbf{I} - \alpha\beta\mathbf{C}^T\mathbf{C})\hat{\mathbf{f}}_i + \beta\mathbf{H}^T(\mathbf{g} - \mathbf{H}\hat{\mathbf{f}}_i). \tag{3.8}$$

CG may be used to speed up the restoration; this will be discussed further in §3.2.3. Spatial adaptivity may be added to suppress ringing and take account of the edge sensitivity and noise-masking of the HVS by performing the iteration in a weighted space [121]. Introducing a diagonal matrix $\mathbf{R}$ which takes values between 0 and 1 according to some estimate of local

image activity, the iteration is modified to:

$$\hat{\mathbf{f}}_{i+1} = \left(\mathbf{I} - \alpha\beta\mathbf{C}^T(\mathbf{I} - \mathbf{R})\mathbf{C}\right)\hat{\mathbf{f}}_i + \beta\mathbf{H}^T\mathbf{R}(\mathbf{g} - \mathbf{H}\hat{\mathbf{f}}_i), \tag{3.9}$$

the first term of (3.8) is effectively a noise-reducing filter, which can be weighted more (values in $(\mathbf{I} - \mathbf{R})$ near 1) in smooth regions and less at edges, and vice-versa for the second, the restoration term. The weights in $\mathbf{R}$ can be updated based on $\hat{\mathbf{f}}_i$ at each iteration.

It is interesting to note that iterative solutions of some sort form the basis of most image restoration algorithms, and many can be traced back to variations on this basic formulation.

3.2.3 Conjugate Gradients Algorithm

Although gradient descent is a fairly intuitive way to solve a minimisation problem, it suffers from poor convergence; for solving a linear system, the rate of convergence depends on the ratio of the largest and smallest eigenvalues, or the conditioning of the system. An ill-conditioned system will tend to have a very slow convergence to the solution.

The method of conjugate gradients (see e.g. [101, 200]) provides a much more efficient way to solve a large linear system $\mathbf{A}\boldsymbol{x} = \mathbf{y}$, where $\mathbf{A}$ is symmetric positive definite. Fortunately the solutions we are interested in fall into this class. One problem with SD is that it does not keep track of the previous search directions at each iteration, so can oscillate around a solution. CG solves this problem by minimising at each iteration over a *Krylov subspace*, formed from repeated application of the matrix $\mathbf{A}$ to the residual vector.

It can be shown that CG will converge to the exact solution of a system of n equations in exactly n iterations (at this point the Krylov subspace spans the whole dimension of the matrix). However for image processing problems n is still too large, and we typically terminate the method early; this again avoids over-converging to noise in the system. Fortunately for typical problems, very good convergence is achieved after just a few hundred iterations. We will make particular use of a variant of CG, termed conjugate gradients least sqaures (CGLS) that operates on the normal equations. For instance we can rewrite (3.5) in the form $\mathbf{M}^T\mathbf{M}\boldsymbol{x} = \mathbf{M}^T\mathbf{y}$, and at each iteration of CGLS we only have to multiply once by $\mathbf{M}^T$ and $\mathbf{M}$ — we don't need to store their product, which is useful to save memory.

3.2.4 Other methods

There are literally hundreds of other methods for the solution of inverse problems, however many are variations on a theme. Some of the other notable techniques used in image restoration will now be summarised. Further methods that have appeared in the literature will be mentioned in §4.4.

3.2.4.1 Projection onto convex sets (POCS)

Non-linear deterministic constraints (usually representing some physical condition) may also be imposed by assuming that the solution is a member of a given closed *convex set*, $\mathcal{S}$, defined in the image space, again representing reasonable constraints on the solution. At each iteration the new estimate is also projected onto the set (i.e. if the solution is not already a member of the set, the closest solution fulfilling the extra constraints is chosen). For example, the amplitude of each pixel may be constrained to a finite range (e.g. positivity) to reduce ringing [121] (see fig. 3.3(b)).

3.2.4.2 Preconditioning

The iterative algorithms may be sped up further by use of *preconditioners* [18, 46, 47, 158], which are basically a method of improving the efficiency of the CG method by premultiplying the LS equation (3.2) with a linear operator that approximates the inverse of the blur, but is well conditioned. If the singular values of the new preconditioned conjugate gradients (PCG) system are clustered closer to 1 then convergence will be more rapid.

3.2.5 Figures of merit and image quality assessment

In order to asses the performance of restoration methods, quality metrics are usually used in addition to visual assessment. The standard measure in deblurring is the improvement in signal-to-noise ratio (ISNR), defined as

$$10 \log_{10} \left(\|\mathbf{f} - \mathbf{g}\|^2 / \|\mathbf{f} - \hat{\mathbf{f}}\|^2 \right). \tag{3.10}$$

It is unfortunate that this metric continues to be used as a means of assessment of image quality, as in many cases it does not correlate well with the subjective appearance of images

ranked by human observers. This is partly due to psychovisual effects such as activity masking where we are less sensitive to noise around strong gradients, but in smooth regions errors are easily spotted. Moreover, it is not invariant to certain distortions such as adding a constant offset that we would class as having negligible effect on the signal content.

For this reason we will also make use of an improved metric, the structural similarity measure (SSIM), which is described in [223], which uses a notion of structural similarity in the image, and in human trials has been shown to correlate a lot more closely with the perceived quality. It gives a score between zero and one to the image under test, with one being a perfect match to the reference, and zero being completely different. However we still have to be careful when using this to measure blind deconvolution fidelity, as it is not invariant to scale or shift of the signal, which can happen due to ambiguity in the solution, without affecting the true quality of the restoration.

3.2.6 Image restoration example

An example of the improvement over the standard CLS filter (the Wiener filter output is very similar) that may be obtained through use of projection onto convex sets (POCS) with adaptively weighted iterative methods is shown in fig. 3.3, where the image in 3.2(b) has been deblurred.

(a) Original Image, f (b) Blurred + Noisy Image, g

Figure 3.2: *Test images, blurred with 4 pixel radius blur at 40dB BSNR*

(a) $\hat{\mathbf{f}}$, image deblurred using CLS filter

(b) $\hat{\mathbf{f}}$, image deblurred using adaptively weighted iterative filter with POCS

Figure 3.3: *Example of deconvolution of fig. 3.2(b) with different methods (BSNR=40dB)*

3.3 Related Inverse problems in Image Processing

A variety of other problems in image processing also fall under the categories of of ill-posed or ill-conditioned inverse problems. As such, they may often be solved using similar techniques to those we will investigate for BID, with an appropriate change in the observation model or objective function. A couple of examples are summarised here; see also the discussion on future work in §11.2.

Denoising The denoising problem, in the case of additive noise is also handled as a special case of non-blind deconvolution, by setting $\mathbf{H} = \mathbf{I}$. Other types of noise, such as impulsive, Poisson, or multiplicative, are often dealt with and require suitable changes to the observation model. It is interesting to note that many of the image priors that have been applied in the literature for image restoration and BID first originated in noise removal applications — perhaps understandable because denoising is less ill-posed, and is therefore a much easier problem to test out new techniques with. Many other heuristic approaches have been proposed though, that don't fit into the framework that we will present.

Inpainting Inpainting is a closely related problem to denoising, as it essentially involves removing a pixelwise degradation. In this case however, some of the pixels are considered

missing and must be estimated based upon the surrounding image (it is usually assumed a mask is given to show which pixels are corrupted).

Image segmentation In segmentation, the goal is to try to assign a class label to each pixel in the image, such that the classes share some properties like adjacency or similar statistical or spectral properties. This label field is essentially hidden data that segmentation algorithms should reconstruct, and thus also falls into the category of inverse problems as there are many segmentations that could have yielded the same observation. Typically similar MRF priors to those used in image restoration are used.

Motion estimation Motion estimation is the problem of finding correspondences between multiple views of the same scene, for instance different frames of a video. This is also an ill-posed problem as there can be several possible matches, with problems such as occlusions at object boundaries. Again, regularisation is a way to provide local consistency to the estimation.

Superresolution If there are several different degraded versions of the image available, that have some geometric distortion or warping between them, they can be fused together through the SR process (see e.g. [112, 174] for reviews). One way to view this problem is non-uniform interpolation of the samples from the various images (the fact that it is possible at all is due to subpixel displacements between the samples). However the full problem also typically requires also simultaneous motion estimation, and possibly also blind deconvolution (BD). It is generally better to consider the joint solution of these problems rather than a sequential approach. The combination of BID and SR has been recently studied in [204].

3.4 Conclusions

In this chapter some background on inverse problems has been given, including discussion of ill-posed problems and their discrete ill-conditioned counterparts. Classical regularisation methods were summarised, and approximate and iterative techniques for the solution of the large problems that are typically faced in image processing were described. In the next chapter, after reviewing some Bayesian theory, these classical regularisation methods are re-derived from a Bayesian perspective.

Chapter 4
Bayesian framework & Prior modelling

In this chapter, we begin in §4.1 with an overview of the Bayesian methodology for inference in inverse problems such as BID. We discuss the probabilistic model of the observation process, and using priors and hyperpriors in the hierarchical framework. We then discuss marginalisation and the use of conjugate priors. In §4.2 we show an example of Bayesian inference for standard image restoration. The most important part of Bayesian modelling for inverse problems is typically the definition of good prior models. In §4.3 we review some models that are used just for blurs, and in §4.4 some models for images (and also blurs) that are used in BID.

4.1 Bayesian Framework

In the methods developed in this thesis, we will be primarily concerned with use of Bayesian inference methods. Bayesian methods offer a flexible and consistent methodology to deal with uncertainty in inference when limited amounts of data are available. As such, they represent an ideal way to tackle ill-posed problems such as BID. An overview of the principles of Bayesian modelling will now be given. In depth discussion and background on Bayesian methods may be found in for example [17] or [73].

4.1.1 Bayesian modelling

A fundamental principle of the Bayesian philosophy is to regard all parameters and observable variables as unknown stochastic quantities, assigning probability distributions based on subjective beliefs. These probability density functions (PDFs) describe the plausibility of a variable taking a particular value, conditional on existing knowledge.

We will use the notation θ to denote a particular unknown variable, and $\boldsymbol{\Theta}$ for the set of all

unknown parameters in our model. Then we may state *Bayes' theorem* which describes the combination of our prior knowledge with the information received from an observation:

$$\underbrace{\mathrm{p}\left(\boldsymbol{\Theta} \,|\, \mathcal{D}, \mathcal{M}\right)}_{\text{posterior}} = \frac{\overbrace{\mathrm{p}\left(\mathcal{D} \,|\, \boldsymbol{\Theta}, \mathcal{M}\right)}^{\text{likelihood}} \overbrace{\mathrm{p}\left(\boldsymbol{\Theta} \,|\, \mathcal{M}\right)}^{\text{prior}}}{\underbrace{\mathrm{p}\left(\mathcal{D} \,|\, \mathcal{M}\right)}_{\text{evidence}}}, \tag{4.1}$$

where $\mathcal{D}$ denotes the observed data, and $\mathcal{M}$ the choice of model (usually this is implicit and only shown when we are choosing between different models). The relationship in Bayes' theorem is obtained by applying simple rules of conditional probability. The distributions have special names as shown in (4.1). However it is the meaning that is important; we are interested in finding the likely values of $\boldsymbol{\Theta}$. The *likelihood*, as a function of $\boldsymbol{\Theta}$, describes the probability of the observation having been generated from the model with parameters $\boldsymbol{\Theta}$. The *prior* describes the existing knowledge about the parameters, before the observation takes place. The *posterior* represents the fusion of this information, and is the distribution we are interested in obtaining and using to infer the values of the unknown parameters. The evidence normally may be regarded as a normalising constant (except when dealing with model selection).

Thus in BID (letting $\boldsymbol{\Theta} = \{\mathbf{f}, \mathbf{h}, \boldsymbol{w}\}$ and $\mathcal{D} = \mathbf{g}$) the original image $\mathbf{f}$, the blur $\mathbf{h}$, and the noise $\boldsymbol{w}$ in Equation (2.6) are all treated as samples of random fields, with corresponding prior PDFs that model our knowledge about the imaging process and the nature of images. These prior distributions, written as $\mathrm{p}\left(\mathbf{f} \,|\, \Omega\right)$, $\mathrm{p}\left(\mathbf{h} \,|\, \Omega\right)$, and $\mathrm{p}\left(\boldsymbol{w} \,|\, \Omega\right)$, depend on parameters which will be denoted by Ω. The parameters of the prior distributions are termed *hyperparameters*.

4.1.2 Bayesian observation model for BID

The first stage of the Bayesian formulation is specifying the likelihood of the observed image, $\mathbf{g}$. In the context of the parameters in BID, the posterior becomes

$$\mathrm{p}\left(\mathbf{f}, \mathbf{h}, \boldsymbol{w} \,|\, \mathbf{g}, \Omega\right) = \frac{\mathrm{p}\left(\mathbf{g} \,|\, \mathbf{f}, \mathbf{h}, \boldsymbol{w}, \Omega\right) \mathrm{p}\left(\mathbf{f}, \mathbf{h}, \boldsymbol{w} \,|\, \Omega\right)}{\mathrm{p}\left(\mathbf{g}\right)}. \tag{4.2}$$

In this most general sense, the observation $\mathbf{g}$ is conditioned on all the other unknowns in the likelihood. However adopting the addtitive noise model and the linear relation $\mathbf{g} = \mathbf{H}\mathbf{f} + \boldsymbol{w}$ from (2.6), there is a *deterministic* relationship between $\mathbf{g}$ and $\boldsymbol{w}$, conditioned on the other

variables $\mathbf{f}$ and $\mathbf{h}$. Therefore the dependence on $\boldsymbol{w}$ is dropped, [1] and the likelihood is instead found via a probability transformation of the PDF of the noise itself.

A typically used model for $\boldsymbol{w}$ is zero mean independent white Gaussian noise (WGN) with distribution $\mathrm{p}_{W}(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w} \,|\, 0, \mathbf{Q}_{w})$, where $\mathbf{Q}_{w}^{-1}$ is a diagonal weights matrix, parameterised by $\boldsymbol{\sigma}_{w}$, included for generality. It allows us to represent spatially varying noise statistics. For stationary noise, $\mathbf{Q}_{w} = \sigma_{w}^{2}\mathbf{I}$, where σ_{w}^{2} is the noise variance. Thus the likelihood of $\mathbf{g}$ conditioned on $\mathbf{h}$, $\mathbf{f}$ and Ω is given by

$$\mathrm{p}_{\mathrm{G}}(\mathbf{g} \,|\, \mathbf{f}, \mathbf{h}, \sigma_{w}) = \mathrm{p}_{W}(\mathbf{g} - \mathbf{H}\mathbf{f}) = \sqrt{\frac{\det|\mathbf{Q}_{w}|^{-1}}{(2\pi)^{\mathrm{L_g}}}} \exp\left[-\frac{1}{2}\|\mathbf{g} - \mathbf{H}\mathbf{f}\|^{2}_{\mathbf{Q}_{w}^{-1}}\right], \tag{4.3}$$

where N is the size of the image vector, $\mathbf{f}$, and the weighted norm used in the PDF is defined as $\|\mathbf{x}\|^{2}_{\mathbf{M}} = \mathbf{x}^{\mathsf{T}}\mathbf{M}\mathbf{x}$ and $\mathbf{M}$ is symmetric positive definite.

4.1.3 Prior modelling

The next stage in specifying the posterior is choice of the prior distributions. The various blur and image priors that have been used for BID will be investigated in §4.3 and 4.4

The "correct" choice of prior distributions is one of the most contentious points of the Bayesian philosophy, according to those who hold a frequentist view of probability. That is, we are assigning a subjective viewpoint to the inference process through the choice of the prior distributions. However this is arguably its most powerful point: enabling us to impart extra expert knowledge to the problem in this fashion, in a principled manner.

Though it is therein that the skill or art of the Bayesian modeller lies. Typically we need to choose carefully the prior distributions that best model the types of signals or data we are trying to estimate — in this case images and blurs. This must be balanced with the needs of tractibility, such that the models do not become too complicated and impractical to perform inference with. Thus traditionally a range of standard probability distributions, such as members of the exponential family (Gaussian, Gamma, Poisson, etc.) are used. This often enables analytic tractibility.

Recent numerical and approximate Bayesian methods, such as MCMC and VB, which will be

[1] Technically, the deterministic relation is represented as $\mathrm{p}(\mathbf{g} \,|\, \mathbf{f}, \mathbf{h}, \boldsymbol{w}, \Omega) = \delta(\boldsymbol{w}, \mathbf{g} - \mathbf{H}\mathbf{f})$, and $\boldsymbol{w}$ is removed by marginalisation (see §4.1.5), as $\mathrm{p}(\mathbf{g} \,|\, \mathbf{f}, \mathbf{h}, \Omega) = \int \mathrm{p}(\mathbf{g} \,|\, \mathbf{f}, \mathbf{h}, \boldsymbol{w}, \Omega)\, \mathrm{p}(\boldsymbol{w})\, \mathrm{d}\boldsymbol{w} = \mathrm{p}_{W}(\boldsymbol{w})|_{\boldsymbol{w}=\mathbf{g}-\mathbf{H}\mathbf{f}}$.

applied in this thesis, do make it easier to work with more general classes of distributions so some of these constraints are no longer necessary and better models may be obtained. More complicated models are also possible by building up from simpler distributions in a hierarchical fashion. This has certain advantages in keeping inference simple, as we will see.

4.1.4 Hierarchical Bayesian Framework

Often, under a standard Bayesian analysis the modelling is complete at this point; the hyperparameters Ω are assumed known (or they are first estimated separately from initial estimates of $\mathbf{f}$ and $\mathbf{h}$).

Alternatively, we may adopt the *hierarchical* Bayesian framework whereby Ω is also assumed unknown, in which case we also model our prior knowledge of its values. The PDFs for the hyperparameters are termed *hyperprior* distributions. This abstraction allows greater robustness to error when there is uncertainty, and is essential when we are less confident in the observed data (due to a lower signal-to-noise ratio (SNR)).

It is typically assumed that $\mathbf{f}$ and $\mathbf{h}$ are *a priori* conditionally independent, given Ω, i.e., $\mathrm{p}(\mathbf{f}, \mathbf{h} \,|\, \Omega) = \mathrm{p}(\mathbf{f} \,|\, \Omega)\, \mathrm{p}(\mathbf{h} \,|\, \Omega)$. Then the task is then to perform inference using the posterior

$$\mathrm{p}(\mathbf{f}, \mathbf{h}, \Omega \,|\, \mathbf{g}) = \frac{\mathrm{p}(\mathbf{g} \,|\, \mathbf{f}, \mathbf{h}, \Omega)\, \mathrm{p}(\mathbf{f} \,|\, \Omega)\, \mathrm{p}(\mathbf{h} \,|\, \Omega)\, \mathrm{p}(\Omega)}{\mathrm{p}(\mathbf{g})}. \tag{4.4}$$

As most Bayesian inference methods (see §5.2) only require the posterior to be known up to a constant of proportionality, we can equally work with the joint global distribution $\mathrm{p}(\Omega, \mathbf{f}, \mathbf{h}, \mathbf{g})$, that is, just the numerator of (4.4).

The hierarchical approach may be continued to further stages as required, such that the *hyper-hyperparameters* that parameterise the hyperpriors are also considered unknown, and so on. This type of modelling transfers uncertainty into specification of the hyperparameters' values into specification of the structure of the model itself, and the inference algorithms become less sensitive to incorrect choices of hyperparameter values. This builds robustness into the inference procedure.

4.1.5 Marginalisation

After modelling the likelihood and the priors that describe the problem, and arriving at the posterior distribution, the next step is to actually use it to calculate the distribution or some

statistics of the unknown parameters. Often this is the most difficult part of the process, because it requires solving a difficult optimisation problem, or evaluating an integral in high dimensions. Therefore various methods have been devised to help find these solutions, and will be discussed in §5.2.

However, often we will only be interested in the value of a subset of the parameters $\boldsymbol{\Theta}$. Bayesian theory provides another important tool, that of *marginalisation*. The principle is that by integrating out the other variables, i.e. taking into account all possible values they might take on, we are left with a distribution containing only the variables of interest. The variables that are integrated over are termed *nuisance parameters* because their estimation is not important at the end of the day. These methods are discussed further in the context of BID in §D.2.

In terms of the other categories of BID methods, we can marginalise either $\mathbf{f}$, to describe an *a priori* blur identification method as

$$\mathrm{p}(\mathbf{h}|\mathbf{g}) = \int \cdots \int \mathrm{p}\left(\mathbf{f}, \mathbf{h}, \Omega \,|\, \mathbf{g}\right) \mathrm{d}\mathbf{f} \cdot \mathrm{d}\Omega, \tag{4.5}$$

or marginalise $\mathbf{h}$ to obtain a *direct* method:

$$\mathrm{p}(\mathbf{f}|\mathbf{g}) = \int \cdots \int \mathrm{p}\left(\mathbf{f}, \mathbf{h}, \Omega \,|\, \mathbf{g}\right) \mathrm{d}\mathbf{h} \cdot \mathrm{d}\Omega. \tag{4.6}$$

These different categories of BID methods are discussed further in §5.1.

4.1.6 Conjugate Priors

An important problem is the estimation of the vector of parameters Ω when they are unknown. To deal with this estimation problem, the hierarchical Bayesian paradigm introduces a second stage (the first stage consisted of the formulation of $\mathrm{p}(\mathbf{f}|\Omega)$, $\mathrm{p}(\mathbf{h}|\Omega)$, and $\mathrm{p}(\mathbf{g}|\mathbf{f}, \mathbf{h}, \Omega)$). In this stage the hyperprior $\mathrm{p}(\Omega)$ is also formulated.

A large part of the Bayesian literature is devoted to finding hyperprior distributions $\mathrm{p}(\Omega)$ for which $\mathrm{p}(\Omega, \mathbf{f}, \mathbf{h}|\mathbf{g})$ can be calculated in a straightforward way or be approximated. These are the so-called conjugate priors [16], which were developed extensively in Raiffa and Schlaifer [180].

Besides providing for easy calculation or approximations of $\mathrm{p}(\Omega, \mathbf{f}, \mathbf{h}|\mathbf{g})$, conjugate priors have,

as we will see later, the intuitive feature of allowing one to begin with a certain functional form for the prior and end up with a posterior of the same functional form, but with the parameters updated by the sample information.

Taking the above considerations about conjugate priors into account, the literature in BID uses different *a priori* models for the parameters depending on the type of unknowns. For parameters, ω, corresponding to inverses of variances, the gamma distribution is used. This is defined by:

$$p(\omega) = \mathcal{GA}\left(\omega \mid \alpha_\omega, \beta_\omega\right) = \frac{(\beta_\omega)^{\alpha_\omega}}{\Gamma(\alpha_\omega)} \omega^{\alpha_\omega - 1} \exp[-\beta_\omega\, \omega], \tag{4.7}$$

where $\omega > 0$ denotes a hyperparameter, $\beta_\omega > 0$ is the scale parameter, and $\alpha_\omega > 0$ is the shape parameter. These parameters are assumed known. The gamma distribution has the following mean, variance, and mode:

$$E[\omega] = \frac{\alpha_\omega}{\beta_\omega}\,, \qquad Var[\omega] = \frac{\alpha_\omega}{(\beta_\omega)^2}\,, \qquad Mode[\omega] = \frac{\alpha_\omega - 1}{\beta_\omega}. \tag{4.8}$$

Note that the mode does not exist when $\alpha_\omega^o \leqslant 1$ and that mean and mode do not coincide.

Equivalently (and we will do so in this thesis), we can use an inverse-Gamma (IG) distribution to model the variance of a Gaussian distribution itself. The IG, defined for positive σ^2 is

$$\mathcal{IG}\left(\sigma^2 \mid \alpha, \beta\right) = \frac{\beta^\alpha}{\Gamma(\alpha)} \sigma^{-2(\alpha+1)} e^{-\frac{\beta}{\sigma^2}}, \tag{4.9}$$

$$\text{with} \qquad \mathbb{E}\left[\sigma^2\right] = \frac{\beta}{\alpha - 1}, \qquad \text{Var}\left[\sigma^2\right] = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)} \tag{4.10}$$

For components of mean vectors the corresponding conjugate prior is a normal distribution. Additionally, for covariance matrices the hyperprior is given by an inverse Wishart distribution (see [73]).

We observe, however, that in general most of the methods proposed in the literature use the *uninformative* prior model

$$p(\Omega) = \text{constant}. \tag{4.11}$$

This is also termed an *improper* prior because it may not be normalised. However it may also be obtained as a limiting case from the standard conjugate priors, for instance by letting the variance of the Gaussian distribution tend to infinity.

4.1.7 Graphical Models

Because we will end up with a lot of variables in the hierarchical models, we will make use of *graphical models* in order to illustrate the relationships between the variables of interest in a concise visual manner. For hierarchal models, these usually take the form of a tree structure; when there are no loops, they are termed directed acyclic graphs (DAGs).

The observation model already presented can be expressed as the root of the graph, as in fig. 4.1. In Fig. 4.1(a), a non-blind restoration problem is shown. Shading of a node indicates the variable is observed or known a-priori. In Fig. 4.1(b), prior models are added to the unknown variables in the blind problem.

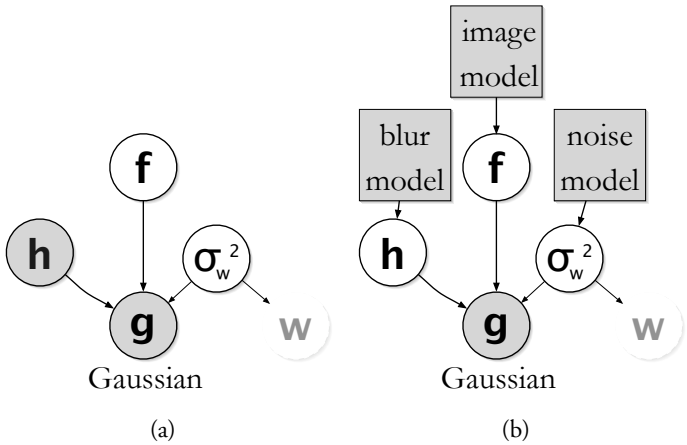


Figure 4.1: *Graphical models showing the relationships between the variables*

4.2 Bayesian image restoration example

The standard Wiener or CLS solutions (§3.1.2.2,3.1.2.1) to the classical image restoration problem may also be obtained via the Bayesian framework. Beginning with the likelihood from (4.3), we may write:

$$p(\mathbf{g} \mid \mathbf{f}, \mathbf{H}, \sigma_w) = \mathcal{N}(\mathbf{g} \mid \mathbf{H}\mathbf{f}, \mathbf{Q}_w) \tag{4.12}$$

$$= \exp\Big(-\frac{1}{2}\big[\mathbf{f}^T(\mathbf{H}^T\mathbf{Q}_w^{-1}\mathbf{H})\mathbf{f} - 2\mathbf{f}^T(\mathbf{H}^T\mathbf{Q}_w^{-1}\mathbf{g}) + \mathbf{g}^T\mathbf{Q}_w^{-1}\mathbf{g}\big]\Big). \tag{4.13}$$

For now we will assume a Gaussian prior for $\mathbf{f}$, with a known covariance matrix $\mathbf{\Sigma}_\mathbf{f}$:

$$\mathrm{p}\left(\mathbf{f}\,|\,\mathbf{\Sigma}_\mathbf{f}\right) = \mathcal{N}\left(\mathbf{f}\,|\,0, \mathbf{\Sigma}_\mathbf{f}\right) \tag{4.14}$$

$$= \exp\Big(-\frac{1}{2}\big[\mathbf{f}^\mathsf{T}\mathbf{\Sigma}_\mathbf{f}^{-1}\mathbf{f}\big]\Big), \tag{4.15}$$

and the posterior is found as

$$\mathrm{p}\left(\mathbf{f}\,|\,\mathbf{g}\right) \propto \mathrm{p}\left(\mathbf{g}\,|\,\mathbf{f}\right)\mathrm{p}\left(\mathbf{f}\,|\,\mathbf{\Sigma}_\mathbf{f}\right) \tag{4.16}$$

$$\propto \exp\Big(-\frac{1}{2}\big[\mathbf{f}^\mathsf{T}(\mathbf{H}^\mathsf{T}\mathbf{Q}_w^{-1}\mathbf{H} + \mathbf{\Sigma}_\mathbf{f}^{-1})\mathbf{f} - 2\mathbf{f}^\mathsf{T}(\mathbf{H}^\mathsf{T}\mathbf{Q}_w^{-1}\mathbf{g}) + \mathbf{g}^\mathsf{T}\mathbf{Q}_w^{-1}\mathbf{g}\big]\Big) \tag{4.17}$$

which is a Gaussian

$$\mathrm{p}\left(\mathbf{f}\,|\,\mathbf{g}\right) \propto \mathcal{N}\left(\mathbf{f}\,\middle|\,\boldsymbol{\mu}_{\hat{\mathbf{f}}}, \mathbf{\Sigma}_{\hat{\mathbf{f}}}\right) \tag{4.18}$$

$$\propto \exp\Big(-\frac{1}{2}\big[\mathbf{f}^\mathsf{T}\mathbf{\Sigma}_{\hat{\mathbf{f}}}^{-1}\mathbf{f} - 2\mathbf{f}^\mathsf{T}\mathbf{\Sigma}_{\hat{\mathbf{f}}}^{-1}\boldsymbol{\mu}_{\hat{\mathbf{f}}} + \boldsymbol{\mu}_{\hat{\mathbf{f}}}^\mathsf{T}\mathbf{\Sigma}_{\hat{\mathbf{f}}}^{-1}\boldsymbol{\mu}_{\hat{\mathbf{f}}}\big]\Big). \tag{4.19}$$

By comparison of (4.17) and (4.19), the parameters are given as

$$\mathbf{\Sigma}_{\hat{\mathbf{f}}}^{-1} = \mathbf{H}^\mathsf{T}\mathbf{Q}_w^{-1}\mathbf{H} + \mathbf{\Sigma}_\mathbf{f}^{-1} \tag{4.20}$$

$$\boldsymbol{\mu}_{\hat{\mathbf{f}}} = \mathbf{\Sigma}_{\hat{\mathbf{f}}}\left(\mathbf{H}^\mathsf{T}\mathbf{Q}_w^{-1}\mathbf{g}\right). \tag{4.21}$$

The mean of this distribution, which is also the maximum, is just

$$\hat{\mathbf{f}} = \boldsymbol{\mu}_{\hat{\mathbf{f}}} = \left(\mathbf{H}^\mathsf{T}\mathbf{Q}_w^{-1}\mathbf{H} + \mathbf{\Sigma}_\mathbf{f}^{-1}\right)^{-1}\left(\mathbf{H}^\mathsf{T}\mathbf{Q}_w^{-1}\mathbf{g}\right) \tag{4.22}$$

$$= \left(\mathbf{H}^\mathsf{T}\mathbf{H} + \sigma_w^2\mathbf{\Sigma}_\mathbf{f}^{-1}\right)^{-1}\mathbf{H}^\mathsf{T}\mathbf{g}, \tag{4.23}$$

where we have substituted a stationary noise covariance, $\mathbf{Q}_w = \sigma_w^2\mathbf{I}$ in the last line.

We now take a look at the types of priors that are used in practice to model the image and blur. The prior distributions $\mathrm{p}(\mathbf{f}|\Omega)$ and $\mathrm{p}(\mathbf{h}|\Omega)$ should reflect our beliefs about the nature of $\mathbf{f}$ and $\mathbf{h}$ and constrain the space of possible solutions for them to the most probable ones. This is necessary due to the ill-posed nature of the problem. Abstract descriptions of natural images have been made: smooth, piecewise-smooth, or textured, for instance (of course some applications may have other specific constraints). Typically this is done in a stochastic sense by specifying probabilistic relations between neighbouring image pixels or their derivatives.

Similar procedures may be followed for the PSF.

4.3 Blur modelling

4.3.1 Types of blur models

Blur models may be categorised as either *parametric* or *non-parametric*. Parametric forms (see §4.3.2) assume the general form of the blur is known, and represent it by a small number of parameters which must be estimated; the PSF $\mathbf{h}$ itself is generated as a deterministic function of these parameters, denoted $\boldsymbol{\theta}_h$. The non-parametric forms (see §4.3.4) (we can also term these stochastic models) simply estimate the coefficients of the PSF directly, that is we specify a prior on their values $p(\mathbf{h})$.

4.3.1.1 Inverse Models

Another categorisation is according to whether the model represents the PSF directly, or models its inverse (approximately) as a restoration filter. The inverse models have the advantage that they can directly filter the blurred image using a finite impulse response (FIR) filter; in their estimation there will be an implicit form of regularisation such that the exact inverse is not estimated, but rather one that produces a good result according to the image model or optimality criteria. The disadvantage of these methods is that they do not give an estimate of the PSF itself. Though they are less common, there are methods in the literature such as [41, 77, 117, 225, 231] that make use of inverse or restoration filter models.

An example of a (parametric) inverse blur model will be used in Chapter 7, because of the lack of regularisation in the formulation, whilst the more complete models developed later in the thesis will use a regular non-parametric PSF model, because regularisation for restoration is provided by the image model.

4.3.2 Parametric Models

In the parametric case, the prior should be placed on the blur model parameters $\boldsymbol{\theta}_h$ themselves; that is we are estimating $\boldsymbol{\theta}_h$ rather than $\mathbf{h}$ itself. Traditionally this prior, $p(\boldsymbol{\theta}_h)$, is assumed to be a uniform distribution however. Then the unknown parameters may be estimated using, for example, maximum likelihood (ML) methods (see [107, 123]). Alternatively the unknown

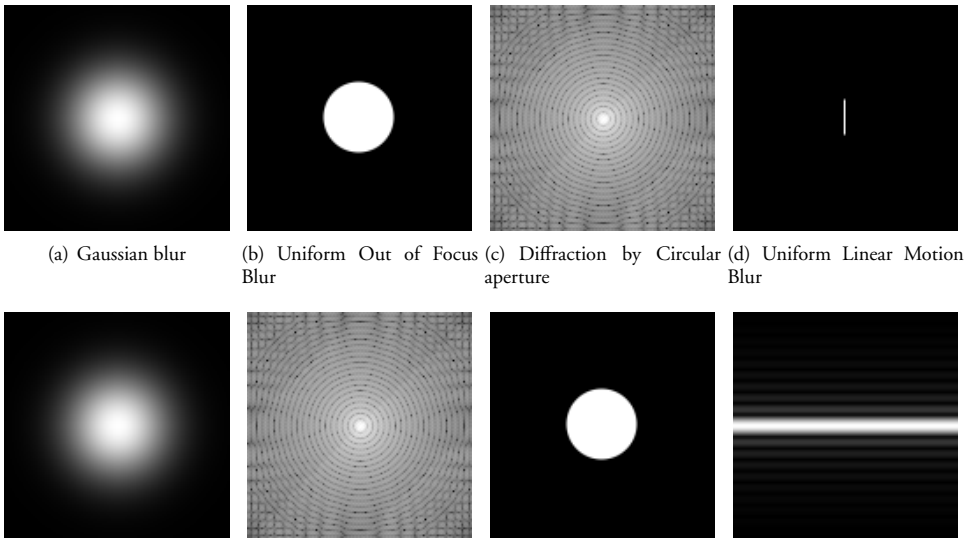
quantities defining the parametric function may be estimated *a priori* if the real underlying image is known. For instance, in astronomy the parameters of the atmospheric turbulence blur are estimated using observed known point sources (stars); similarly, the parameters describing the HST PSF model could also be estimated.

The first BID methods (see §C.1) concentrated on identifying features of the spectra of these parametric models in the blurred observation; for instance the zero crossings of the Bessel function type spectra result in an image defocused through a circular aperture.

Parametric models are often too constrained to accurately represent all of the effects that have been described in §A.1. However some approaches that have appeared in the literature or are commonly used in simulations are described in the next section.

4.3.3 Example parametric blur models

Some examples of single-parameter parametric blur models, along with their Fourier spectra, are shown in fig. 4.2. Some of these analytic blur models are now summarised.



(a) Gaussian blur (b) Uniform Out of Focus Blur (c) Diffraction by Circular aperture (d) Uniform Linear Motion Blur

(e) Gaussian blur (f) Uniform Out of Focus Blur (g) Diffraction by Circular aperture (h) Uniform Linear Motion Blur

Figure 4.2: *Some simple parametric blur models (a-d) and their spectra (e-h)*

4.3.3.1 Atmospheric Turbulence Blur

This type of blur is common in remote sensing and aerial imaging applications. For long-term exposure through the atmosphere a Gaussian PSF model is used:

$$h(\mathbf{r}) = K e^{-\frac{|\mathbf{r}|^2}{2\sigma^2}}, \tag{4.24}$$

where K is a normalizing constant ensuring that the blur has a unit volume, and σ^2 is the variance that determines the severity of the blur. Alternative atmospheric blur models have been suggested in [146, 147]. In these works the PSF is approximated by the function

$$h(\mathbf{r}) \propto (1 + \frac{|\mathbf{r}|^2}{R^2})^{-\delta}, \tag{4.25}$$

where δ and R are unknown parameters. Gaussian blur is also used as a model for blur in X-ray images, and sometimes for sensor blur, or in other general cases where there is no particular information as to the shape of the blur.

4.3.3.2 General Motion Blur

In general, relative motion of the camera and scene to be imaged results in a PSF representing temporal integration along this motion path. If the motion path is known, or may be estimated through some other means, a PSF model may be found. In general this will be non-stationary, especially for object movement within the scene, and the PSF expression is found by integrating along the path, $\mathbf{p_s}(t) = [p_{x_1;\mathbf{s}}(t), p_{x_2;\mathbf{s}}(t)]^T$, of each pixel $\mathbf{s}$ of the true image as a function of time, $0 \leqslant t \leqslant T$, of the exposure interval:

$$h(\mathbf{x};\mathbf{s}) = \frac{1}{T}\int_0^T \delta\left(x_1 - p_{x_1;\mathbf{s}}(t), x_2 - p_{x_2;\mathbf{s}}(t)\right) dt \tag{4.26}$$

In practice such a motion path could only be estimated if multiple images are available [203], or a secondary device operating at a higher frame rate is available [15].

4.3.3.3 Linear Motion Blur

If the camera movement or object motion is fast relative to the exposure period, we may approximate this as a linear motion blur. This is represented as the 1-D local averaging of

neighbouring pixels. An example of a horizontal motion blur model is given by (L is an even integer):

$$h(\mathbf{r}) = \begin{cases} \dfrac{1}{L+1}, & -\dfrac{L}{2} \leqslant r_1 \leqslant \dfrac{L}{2}, \\ & r_2 = 0 \\ 0 \quad , & \text{otherwise.} \end{cases} \tag{4.27}$$

4.3.3.4 Out-of-Focus Blur

Photographic defocusing is another common type of blurring, as discussed in §A.5. A complete model could represent the various effects due to lens aberrations and aperture shape. However accurate knowledge of all of the parameters involved in the optical system is usually not available after the picture was taken. Thus when the blur due to defocusing is large, the uniform circular PSF model is often used as an approximation to these effects:

$$h(\mathbf{r}) = \begin{cases} \dfrac{1}{\pi r^2}, & |\mathbf{r}| \leqslant r \\ 0 \quad , & \text{otherwise.} \end{cases} \tag{4.28}$$

The frequency response of this PSF is given by a first order Bessel function:

$$\mathcal{H}(\boldsymbol{\omega}) = \frac{J_1(r\,|\boldsymbol{\omega}|^2)}{r\,|\boldsymbol{\omega}|^2} \tag{4.29}$$

This has a radially periodic pattern of zeros, as can be seen in fig. 4.2, and hence the model has been used in the methods in §C.1.

The uniform 2-D blur (square PSF) is sometimes used as a cruder approximation to an out-of-focus blur; it is also used as a model for sensor pixel integration in SR restoration. This model is defined (with L an even integer) as

$$h(\mathbf{r}) = \begin{cases} \dfrac{1}{(L+1)^2}, & -\dfrac{L}{2} \leqslant (r_1, r_2) \leqslant \dfrac{L}{2} \\ 0 \quad , & \text{otherwise.} \end{cases} \tag{4.30}$$

4.3.3.5 Diffraction

Diffraction due to the camera aperture is discussed in §A.9. It is modelled simply, when it is the dominating effect, as the Fourier transform (FT) of the aperture shape. Thus for a circular

aperture the parametric PSF is simply a Bessel function, as described above for the frequency response of the circular PSF.

4.3.4 Non-Parametric Models

Describing a prior that models general stochastic properties of the PSFs of interest is not a trivial task. There are physical constraints arising from the imaging system that should ideally be reflected in any blur prior: positivity of the coefficients (light cannot have negative intensity) and unit gain at zero frequency (the PSF coefficients should sum to one). In terms of the frequency response, optical systems tend to spread out energy and thus have generally low-pass responses. Other prior assumptions often used for the PSF coefficients are smoothness or piecewise smoothness, symmetry, and finite support size. Often in non-Bayesian methods these constraints are enforced using deterministic constraints such as POCS.

In fact, the blur modelling problem has not seen much specific attention in the literature. It has not been a focus of this work to develop new blur priors as such; however some suggestions are made in the future work section §11.2. Instead, the non-parametric models that will be used will be of the same form as those used for the image model, such as those described in §4.4 (which will not be repeated here), or one of the new models developed in later chapters. It may be questioned if these are so well suited for modelling a PSF which is a fundamentally different quantity to a regular image; however at the abstract level we may observe several of the common parametric models possess features such as smooth regions or edges (in the case of a pillbox blur). Thus priors on the gradients or other features of the PSF are one viable option.

4.3.5 Extensions to simple parametric models

Overall, the basic parametric models are only good as a first approximation. We will mainly be interested in modelling defocusing blurs, simple planar motion blurs, and blurs due to general optical imprecision or sensor integration. The sensor response has often been ignored (apart from simple models used in superresolution methods). Real world examples of PSFs from defocused images[2] may indicate the true blur is somewhere in between a Gaussian and a

[2] In the photography world, the aesthetic quality of the defocus is known as *Bokeh* — bad Bokeh may manifest as enhanced edges of the circle of confusion (CoC); more technically the photographic blur is described by the optical transfer function (OTF) and modulation transfer function (MTF), i.e. the normalised frequency response and its magnitude.

circular blur (it will generally be a low pass function). Symmetricity is usually an approximately valid assumption (at least near to the axis of a lens system, ignoring other aberrations).

It would be useful to have a prior model that can include some knowledge of typical optical aberrations found in a typical camera system, and provide a model with a manageable number of parameters to estimate, though as far as is known no such prior has been used in any image restoration task. [239] has provided one idea that using a discrete cosine transform (DCT) decomposition for the blur, i.e. representation using basis functions, reduces parameterisation of the blur due to the assumption that the transform compresses energy, and thus a few coefficients will be enough to provide a good approximation of the PSF. TLS based approaches (§C.6) provide one option for stochastic blur representation. Another interesting "soft" blur model has been proposed by Yap *et al.*[57, 233], whereby the blur is constrained if it is in proximity to a standard model, otherwise it is allowed to deviate further to allow arbitrary blurs.

We may seek to postulate some more appropriate model built of basis functions that are derived from properties of known samples of blurs. It may be necessary as such to consider available samples of MTFs from commercial lens specifications, or to perform measurements to acquire a database of example blurs. Categorisation would enable further hierarchical modelling to be performed; for example PSF profiles may change for the same camera/lens with position away from the axis, focal length, or degree of defocus.

[28, 66] mention some other degradation effects including compression, analogue video distortions, and charge coupled device (CCD) responses, that would be useful to be considered in the context of superresolution, but these are equally applicable to general image restoration. Some work has been done on incorporating, for example, compression artefacts into the restoration problem [196]; images from consumer digital cameras are very likely to have suffered from compression and other pre-processing. Effects of inter-channel blur in coloured images have also received little attention; while there are demosaicing algorithms that attempt to remove Bayer patterns from CCD images, there has been little attempt to combine these with general restoration procedures. Providing an adequate model of the whole degradation process is likely to be an ultimate requirement for improvement to restoration algorithms.

4.4 Image modelling

In this section, we give an overview of the various models that have been used for modelling images, in particular those used for BID. However as these models are also often used as non-parametric blur models, as described in §4.3.4, they will also be described in this context here.

We will consider a general exponential model of the form

$$p(\mathbf{f} \mid \Omega) = \frac{1}{Z_f(\Omega)} \exp\left[-U_f(\mathbf{f}, \Omega)\right] \tag{4.31a}$$

$$p(\mathbf{h} \mid \Omega) = \frac{1}{Z_h(\Omega)} \exp\left[-U_h(\mathbf{h}, \Omega)\right] \tag{4.31b}$$

to represent the image and blur priors. The normalizing terms Z_f and Z_h depend on the hyperparameters for each distribution. They may be treated as constants if we assume the hyperparameters to be known; otherwise they must be calculated as $\int \exp\left[-U_f(\mathbf{f}, \Omega)\right] d\mathbf{f}$ and $\int \exp\left[-U_h(\mathbf{h}, \Omega)\right] d\mathbf{h}$, respectively, which may cause difficulties in inference unless we assume a special form for $U(\cdot)$. Note that $U(\cdot)$ is sometimes termed the *energy function*.

Many different image and blur models in the literature can be put in the form of Equation (4.31); particular cases will now be considered.

4.4.1 Stationary Gaussian Models

The most common model is the class of Gaussian models provided by $U_f = \frac{1}{2}\alpha\|\mathbf{L}\mathbf{f}\|^2$. Then, if $\det|\mathbf{L}| \neq 0$, the term Z_f in Equation (4.31) becomes simply $(2\pi)^{\frac{N}{2}} \alpha^{-\frac{N}{2}} \det|\mathbf{L}|^{-1}$, which if we use a fixed stationary form for $\mathbf{L}$ is simple to calculate. These models are often termed simultaneous autoregression (SAR) or conditional autoregression (CAR) models [189].

In the most basic case, we can use $\mathbf{L} = \mathbf{I}$, the identity. This imposes constraints on the magnitude of the intensity distribution of $\mathbf{f}$. A more common usage is $\mathbf{L} = \mathbf{C}$, the discrete Laplacian operator, which instead constrains the derivative of the image. For instance, Molina et al. [153] used this model for both image and blur, giving

$$p(\mathbf{f} \mid \alpha_{im}) \propto \alpha_{im}^{N/2} \exp\left[-\frac{1}{2}\alpha_{im} \parallel \mathbf{C}\mathbf{f} \parallel^2\right] \tag{4.32a}$$

$$p(\mathbf{h} \mid \alpha_{bl}) \propto \alpha_{bl}^{M/2} \exp\left[-\frac{1}{2}\alpha_{bl} \parallel \mathbf{C}\mathbf{h} \parallel^2\right]. \tag{4.32b}$$

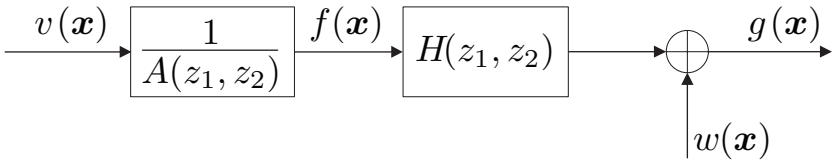


Figure 4.3: *ARMA degradation model.*

Note that in these two equations N and M should in practice be replaced by $N-1$ and $M-1$, respectively, because $\mathbf{C}^T\mathbf{C}$ is singular.

This SAR model is suitable for $\mathbf{f}$ and $\mathbf{h}$ if it is assumed that the luminosity distribution is smooth on the image domain, and that the PSF is a partially smooth function.

In [132] the SAR model was used for the image prior, and a Gaussian PDF for the PSF; that is,

$$p(\mathbf{h} \mid \boldsymbol{\mu}_{bl}, \alpha_{bl}) = \mathcal{N}(\boldsymbol{\mu}_{bl}, \alpha_{bl}^{-1}\mathbf{I}). \tag{4.33}$$

The components of the blur are assumed statistically independent so that the covariance matrix is diagonal. Clearly, a Gaussian PDF with unknown mean vector and fully populated covariance matrix might be used to model image and blur priors. However, this requires the simultaneous estimation of a very large number of hyperparameters, thus making the approach highly impractical unless we use additional hyperpriors (see §4.1.4) to impart *a priori* knowledge of the hyperparameters' values [153].

4.4.2 Autoregressive Models

A class of blind image deconvolution algorithms (see, e.g., [107, 123]) model the observed image $\mathbf{g}$ as an autoregressive moving average (ARMA) process, as shown in Figure 4.3.

The observation (2.6) forms the moving average (MA) part of the model. The original image is modeled as a 2-D autoregressive (AR) process:

$$f(\mathbf{s}) = \sum_{\mathbf{r} \in S_a : \mathbf{r} \neq \mathbf{0}} a(\mathbf{r})\, f(\mathbf{s} - \mathbf{r}) + v(\mathbf{s}), \tag{4.34}$$

or in matrix-vector form:

$$\mathbf{f} = \mathbf{A}\mathbf{f} + \boldsymbol{\nu}, \tag{4.35a}$$

or equivalently,

$$\mathbf{f} = \mathbf{F}\mathbf{a} + \boldsymbol{\nu}, \tag{4.35b}$$

where $S_a \subset \mathbb{R}^2$ is the support of the 2-D AR coefficients $a(\mathbf{r})$, and $\mathbf{A}$ has a BTTB form. The *excitation* noise signal, or modelling error, $\nu(\mathbf{s})$ is a zero-mean WGN process with diagonal covariance matrix $\mathbf{Q}_\nu$, that is independent of $f(\mathbf{s})$. Since $\boldsymbol{\nu} = (\mathbf{I} - \mathbf{A})\mathbf{f}$, the PDF of $\mathbf{f}$ is obtained via a probability transformation of $\boldsymbol{\nu}$, as a Gaussian: $p(\mathbf{f} \mid \mathbf{a}, \mathbf{Q}_\nu) = \mathcal{N}(\mathbf{f} \mid 0, \Sigma_\mathbf{f})$. The image covariance matrix is defined as

$$\Sigma_\mathbf{f} = (\mathbf{I} - \mathbf{A})^{-1} \mathbf{Q}_\nu (\mathbf{I} - \mathbf{A})^{-T}. \tag{4.36}$$

It should be clear that $p(\mathbf{f})$ is in the form of Equation (4.31), with $U_\mathbf{f} = \frac{1}{2}\|(\mathbf{I} - \mathbf{A})\mathbf{f}\|^2_{\mathbf{Q}_\nu}$ and $Z_\mathbf{f} = (2\pi)^{\frac{N}{2}} \det|\mathbf{Q}_\nu|^{\frac{1}{2}} \det|\mathbf{I} - \mathbf{A}|^{-1}$. Unlike the SAR model above, however, where a deterministic form is used for the matrix $\mathbf{L}$, the AR coefficients defining $\mathbf{A}$ and the excitation noise covariance matrix $\mathbf{Q}_\nu$ also have to be estimated (in [123] a flat prior distribution is assumed on them).

A related formulation to the stationary ARMA model is also considered by Katsaggelos and Lay in [107, 108, 128]. In these works, the AR model parameters are not estimated directly, but rather the defining sequence of the matrix $\Sigma_\mathbf{f}$ is found in the discrete frequency domain, along with the other parameters, under the assumption that the image model is stationary. Observe that this approach does not assume a known model support size S_a or blur support S_h.

The AR image model is essentially a higher-order Gaussian MRF model. It has been shown to be good at representing textured images (e.g. [55]); however, like any of the other stationary Gaussian models in this section, it is not such a good model for an original image that has prominent edges as part of the scene. Therefore it is possible to consider nonstationary extensions to the AR model, which can adapt to the local features in the image. This will form the basis of the image models used later in this thesis.

4.4.3 Markov Random Field Models

A class of models encountered extensively in image segmentation [61], classical image restoration [76], and also in superresolution restoration [194] and BID [58, 239] are the Markov random field (MRF) models [226]. They are usually derived using local spatial dependencies; however, we may see that they are closely related to the other models in this section.

Defining $U = \sum_{c \in \mathcal{C}} V_c(\mathbf{f})$ in Equation (4.31), we have the definition of a *Gibbs distribution*. In this context, Z is termed the partition function. $V_c(\mathbf{f})$ is a *potential function* defined over *cliques*, c in the image [226]. Briefly speaking, this gives a simple way of specifying interactions over local neighbourhoods in the image field. If we use quadratic potentials, $V_c(\mathbf{f}) = \left(\mathbf{d}_c^T \mathbf{f}\right)^2$, we may rewrite the Gibbs distribution as a Gaussian:

$$p(\mathbf{f}) = \frac{1}{Z} \exp\left[-\mathbf{f}^T \mathbf{B} \mathbf{f}\right] = \frac{1}{Z} \exp\left[-\sum_{c \in \mathcal{C}} \mathbf{f}^T \mathbf{B}_c \mathbf{f}\right] \tag{4.37}$$

where $\mathbf{B}_c$ is obtained from $\mathbf{d}_c$ and satisfies $[\mathbf{B}_c]_{\mathbf{x},\mathbf{s}}$ are only nonzero when pixels $\mathbf{x}$ and $\mathbf{s}$ are neighbours. Typically the vectors $\mathbf{d}_c$ represent finite difference operators. The partition function is now equal to $(2\pi)^{\frac{N}{2}} \det|\mathbf{B}|^{-\frac{1}{2}}$. This model is also then termed a Gaussian Markov random field (GMRF) [34] or CAR [189].

We may also use generalised Gaussian MRFs (GGMRFs) with arbitrary nonquadratic potentials of a similar functional form: $V_c(\mathbf{f}) = \rho(\mathbf{d}_c^T \mathbf{f})$, where ρ is some (usually convex) function, such as the *Huber function* [34] or p-norm (with $p \geqslant 1$) based function, $\rho(u) = |u|^p$. This is similar to the use of potential functions used in anisotropic diffusion methods, the motivation being edge preservation in the reconstructed image. Other extensions to the model consider hierarchical, or compound GMRFs (CGMRFs), also with the goal of avoiding oversmoothing of edges [76, 99]. We elaborate further on MRF models and summarise a variety of other image models in Appendix B.

4.5 Conclusions

In this chapter we began by introducing the principles of the hierarchical Bayesian framework, and discussed its application to formalising the BID problem. First the observation model was described, based on the uncertainty due to the additive noise, then the need for prior modelling was discussed, and the concepts of hyperpriors and conjugate priors were introduced, as well

as the technique of marginalisation for parameter space reduction.

In the second half of the chapter, we have reviewed some of the existing prior models that have been used for blurs and images in image restoration problems and BID in particular. We have concluded that although parametric modelling is a useful way to reduce the dimension of the coefficients needed to represent the PSF, most of the existing models are somewhat inadequate when we consider the types of degradations that are encountered in practice (see §A.1). Therefore until more work is performed on blur modelling, non-parametric PSFs are the most suitable choice for BID techniques, coupled with known constraints about the blur such as positivity and perhaps piecewise smoothness. In terms of the image models, there are a wealth of choices and different uses in the literature. We focused on models that have been used in BID, and on unifying some of these models using MRF concepts (the connections with variational methods such as total variation (TV) are also discussed in §B.4). In the next chapter, we continue the review with a look at the solution methodologies that have been proposed for BID.

Chapter 5
Solutions to the BID problem

In this chapter, we present an overview of the estimation techniques that can be used to solve the BID problem. In the process of doing so, we survey the literature on BID to categorise and discuss the different approaches that have been used in the past. Many of the techniques fall under the realm of Bayesian inference methods, and we collect these here. Non-Bayesian approaches are reviewd in Appendix C.

5.1 Classification of Blind Image Deconvolution Methodologies

We may classify BID approaches into three categories according to at what stage we identify the blur: *a priori*, jointly with the image, or not at all.

***A priori* blur identification methods** With this approach, the PSF is identified separately from the original image, and later used in combination with one of the classical image restoration algorithms in order to restore the original image. A parametric blur model may be used, for example, one of the general models described in §4.3.2; then the objective is to identify the most likely blur parameters $\mathbf{h}$ from the observation. This approach has been used in [42, 208], for example, and in a Bayesian context in Chapter 7 and [24]. The Bayesian case corresponds to using (4.5) to first estimate the blur, before using this estimate for image restoration.

Experimental approaches are also possible: images of one or more point sources are collected and used to obtain an estimate of the PSF (this, for example, was done with the Hubble space telescope (HST)). Furthermore if a good understanding of the imaging system is available for a specific application, we may make an *a priori* prediction of the blur; however unless we use this model as a prior in one of the other algorithms, this is not a blind procedure as such. This may be possible in microscopy, medical ultrasound, remote sensing, or optical telescope systems (e.g., Tiny Tim modelling of the HST).

Joint Identification methods The majority of existing methods fall into this class, where the image and blur are identified simultaneously. However, in practice many methods in this category use an alternating approach to estimate $\mathbf{f}$ and $\mathbf{h}$ rather than truly finding the joint solution. Prior knowledge about the image and blur is typically incorporated in the form of models like those presented in §4.3.1 and §4.4. Parameters describing such models are also required to be estimated from the available data; often this is performed before image and blur identification, although simultaneous identification is possible, see e.g., [153]. We also estimate all parameters in Chapters 9 and 10.

Direct methods There exists a third class of methods, which seek to bypass the blur identification stage and estimate the image $\mathbf{f}$ directly, by using appropriate constraints or prior knowledge. Although this is less common in practice, it has been done in the multichannel case [170] where extra constraints are available (all the information is there, but "mixed up" between the channels).

Direct image estimation may also be achieved by marginalising $\mathbf{h}$ as in (4.6); this approach does not appear to have been investigated for BID, but was recently applied to the related SR problem by Pickup *et al.*[175, 176]. This work tends to show the superiority of the direct method over an *a priori* approach of first marginalising the image as in [212], which uses a similar approach for SR as will be demonstrated for BID in Chapter 7. There are some differences in the observation model and the type of parameters $\boldsymbol{\theta}$ that are being estimated in the SR case, but in principle if this could be applied to the single channel BID problem then it is likely similar improvements in performance could be obtained.

In §5.2, we use the Bayesian framework to present and categorise the different BID approaches proposed in the literature. As we will see, most of them can be seen as special cases of an application of Bayes' theorem: the main differences between them are the choices of the function to be optimized, and the prior distributions used to model the original image and the degradation process. Using the Bayesian framework allows us to describe the general BID problem in a systematic way, and to identify the similarities and differences between the proposed approaches in the above two categories. Each category may be seen as particular inference model in the Bayesian paradigm.

5.2 Bayesian Inference methods

There are a number of different ways that we may proceed to estimate the image and blur using Equation (4.1). Depending on the prior models chosen, finding analytic solutions may be difficult, so approximations are often needed. Many methods in the literature seek point estimates of the parameters $\mathbf{f}$ and $\mathbf{h}$. Typically, this reduces the problem to one of optimisation. However, the Bayesian framework provides other methodologies for estimating the *distributions* of the parameters [73, 100, 160], which deal better with uncertainty; approximating or simulating the posterior distribution are two options. These different inference strategies and examples of their use will now be presented, proceeding from the simplest to the more complex. There will be an emphasis placed on the VB and MCMC methods, in §5.2.5 and §5.2.6 respectively, that will be used in the second part of this thesis.

5.2.1 Maximum *a Posteriori* and Maximum Likelihood

One possible point estimate is provided by the maximum a posteriori (MAP) solution, which are the values of $\mathbf{f}$, $\mathbf{h}$, and $\boldsymbol{\Omega}$ that maximise the posterior probability density:

$$\{\hat{\mathbf{f}}, \hat{\mathbf{h}}, \hat{\boldsymbol{\Omega}}\}_{\text{MAP}} = \underset{\mathbf{f},\mathbf{h},\boldsymbol{\Omega}}{\operatorname{argmax}} \; \mathrm{p}\left(\mathbf{g} \,|\, \mathbf{f}, \mathbf{h}, \boldsymbol{\Omega}\right) \mathrm{p}\left(\mathbf{f} \,|\, \boldsymbol{\Omega}\right) \mathrm{p}\left(\mathbf{h} \,|\, \boldsymbol{\Omega}\right) \mathrm{p}\left(\boldsymbol{\Omega}\right). \tag{5.1}$$

The *maximum likelihood (ML)* solution attempts instead to maximise the likelihood $\mathrm{p}\left(\mathbf{g} \,|\, \mathbf{f}, \mathbf{h}, \boldsymbol{\Omega}\right)$ with respect to the parameters:

$$\{\hat{\mathbf{f}}, \hat{\mathbf{h}}, \hat{\boldsymbol{\Omega}}\}_{\text{ML}} = \underset{\mathbf{f},\mathbf{h},\boldsymbol{\Omega}}{\operatorname{argmax}} \; \mathrm{p}\left(\mathbf{g} \,|\, \mathbf{f}, \mathbf{h}, \boldsymbol{\Omega}\right). \tag{5.2}$$

Note, however, that in this case we can only estimate the parameters in $\boldsymbol{\Omega}$ that are present in the conditional distribution $\mathrm{p}\left(\mathbf{g} \,|\, \mathbf{f}, \mathbf{h}, \boldsymbol{\Omega}\right)$ but none of those present only in $\mathrm{p}\left(\mathbf{f}, \mathbf{h} \,|\, \boldsymbol{\Omega}\right)$.

The above maximisation of the likelihood is typically seen as a non-Bayesian method, although it is identical to the MAP solution with uninformative (flat) prior distributions. Some approaches may use flat priors for some parameters but not others. Assuming known values for the parameters is equivalent to using degenerate distributions (delta functions) for priors.

For instance, a degenerate distribution on $\boldsymbol{\Omega}$ is defined as

$$\mathrm{p}\left(\boldsymbol{\Omega}\right) = \delta\left(\boldsymbol{\Omega}, \boldsymbol{\Omega}_0\right) = \begin{cases} 1, & \text{if } \boldsymbol{\Omega} = \boldsymbol{\Omega}_0 \\ 0, & \text{otherwise.} \end{cases} \tag{5.3}$$

Then, the *MAP* and *ML* solutions become, respectively,

$$\{\hat{\mathbf{f}}, \hat{\mathbf{h}}\}_{\mathrm{MAP}} = \underset{\mathbf{f},\mathbf{h}}{\operatorname{argmax}} \ \mathrm{p}\left(\mathbf{g} \,|\, \mathbf{f}, \mathbf{h}, \boldsymbol{\Omega}_0\right) \mathrm{p}\left(\mathbf{f} \,|\, \boldsymbol{\Omega}_0\right) \mathrm{p}\left(\mathbf{h} \,|\, \boldsymbol{\Omega}_0\right) \tag{5.4}$$

$$\{\hat{\mathbf{f}}, \hat{\mathbf{h}}\}_{\mathrm{ML}} = \underset{\mathbf{f},\mathbf{h}}{\operatorname{argmax}} \ \mathrm{p}\left(\mathbf{g} \,|\, \mathbf{f}, \mathbf{h}, \boldsymbol{\Omega}_0\right). \tag{5.5}$$

Many deconvolution methods can fit into this Bayesian formulation. The main differences among these algorithms come from the form of the likelihood, the particular choice of priors on the image, blur, and the hyperparameters, and the optimisation methods used to find the solutions.

Observe that the regularization-based approaches using the L_2 norm frequently found in the literature also fall into this category. In these approaches the blind deconvolution problem is stated as a constrained minimization problem, where a cost function is minimized with a number of regularization constraint terms.

In regularization approaches the cost function is chosen as the error function $\|\mathbf{g} - \mathbf{H}\mathbf{f}\|^2_{W}$, which ensures fidelity to the data. The regularization terms are used to impose additional constraints on the optimisation problem. Generally, these constraints ensure smoothness of the image and the blur, that is, the high-frequency energy of the image and the blur is minimized. The effect of the regularization terms is controlled by the regularization parameters, which basically represent the trade-off between fidelity to the data and desirable properties (smoothness) of the solutions.

For example, in [235], the classical regularized image restoration formulation used in [105, 109, 121] was extended to the BID case by adding a constraint for the blur. The problem is stated, in a relaxed minimization form, as

$$\hat{\mathbf{f}}, \hat{\mathbf{h}} = \underset{\mathbf{f},\mathbf{h}}{\operatorname{argmin}} \ \left[\|\mathbf{g} - \mathbf{H}\mathbf{f}\|^2_{\mathbf{Q}_w^{-1}} + \lambda_1 \|\mathbf{L}_{\mathrm{f}}\mathbf{f}\|^2 + \lambda_2 \|\mathbf{L}_{\mathrm{h}}\mathbf{h}\|^2\right], \tag{5.6}$$

where λ_1 and λ_2 are the Lagrange multipliers for each constraint, and $\mathbf{L}_f$ and $\mathbf{L}_h$ are the regularization operators. In [235] each $\mathbf{L}$ is the Laplacian multiplied by a spatially varying weights term, calculated as in [63, 102, 105, 109] from the local image variance in order to provide some spatial adaptivity to avoid oversmoothing edges.

5.2.1.1 Alternating Minimisation, or Iterated Conditional Modes

Let us consider again the solution of Equation (5.4). A major problem in the optimisation is the simultaneous estimation of the variables $\mathbf{f}$ and $\mathbf{h}$. A widely used approach is that of alternating minimization (AM) of Equation (5.6) (or its continuous equivalent for partial differential equation (PDE) formulations), which follows the steepest descent with respect to one unknown while holding the other unknown constant. The advantage of this algorithm is its simplicity due to the linearization of the objective function. This optimisation procedure corresponds to the iterated conditional modes (ICM) proposed by Besag [20].

This estimation procedure has been applied to standard regularization approaches [57, 235], and to the anisotropic diffusion and TV type models described in §B.4, where for example, the objective functional becomes

$$\int_{S_f}\left(g(\mathbf{x})-\int_{S_h} h(\mathbf{s}-\mathbf{x})*f(\mathbf{x})d\mathbf{s}\right)^2 d\mathbf{x}+\lambda_1\int_{S_f}\kappa\left(|\nabla f(\mathbf{x})|\right)d\mathbf{x}+\lambda_2\int_{S_h}\upsilon\left(|\nabla h(\mathbf{s})|\right)d\mathbf{s}. \quad (5.7)$$

Partial derivatives with respect to f and h are taken to give the two PDEs for AM.

There are various numerical methods to solve the associated PDEs. These include the classical Euler, Newton, or Runge–Kutta methods; or recently developed approaches, such as time-marching [191], primal-dual methods [51], lagged diffusivity fixed point schemes [220], and half-quadratic regularization [45] (similar to the discrete schemes in [74, 75]). All of these methods employ techniques to discretize and linearize, or otherwise approximate the solution to the PDEs. The selection of a particular method depends on the computational limitations and speed requirements, since different techniques have different simplicity, stability, and convergence speed properties.

It should be pointed out that although the origination of the TV models is formulated in the continuous domain, these models must be discretised — following which they acquire a similar form to those standard regularisation approaches where the objective function is formed in the discrete domain.

5.2.2 Minimum Mean Squared Error

The MAP estimate does not take into account the whole posterior PDF. If the posterior is sharply peaked about the maximum then this does not matter (typically most probability mass will be contained close to the peak in this case and the two estimates will be similar); however, in the case of high observation noise or a broad (heavy-tailed) posterior this estimate is likely to be unreliable. As mentioned in [155], for a Gaussian in high dimensions most of the probability *mass* is concentrated away from the probability *density* peak.

The minimum mean-squared error (MMSE) estimate attempts to find the optimal parameter values that minimize the expected mean squared error between the estimates and the true values. In other words we aim at calculating the mean value of $\mathrm{p}(\mathbf{f}, \mathbf{h}, \boldsymbol{\Omega}|\mathbf{g})$. In practice, finding MMSE estimates analytically is generally difficult, though it is possible with sampling-based methods (§5.2.6) and can be approximated using VB methods §5.2.5).

5.2.3 Marginalizing Hidden Variables

In the discussion so far none of the unknowns $\mathbf{f}, \mathbf{h}$ and $\boldsymbol{\Omega}$ have been marginalised out to perform inference on only a subset of $\mathbf{f}, \mathbf{h}$, and $\boldsymbol{\Omega}$.

We can, however, approach the BID inference problem by first calculating

$$\hat{\mathbf{h}}, \hat{\boldsymbol{\Omega}} = \underset{\mathbf{h},\boldsymbol{\Omega}}{\operatorname{argmax}} \int_{\mathbf{f}} \mathrm{p}(\mathbf{g}|\boldsymbol{\Omega}, \mathbf{f}, \mathbf{h})\, \mathrm{p}(\mathbf{f}, \mathbf{h}|\boldsymbol{\Omega})\, \mathrm{p}(\boldsymbol{\Omega}) \mathrm{d}\mathbf{f} \tag{5.8}$$

and then selecting as restoration the image

$$\hat{\mathbf{f}}\Big|_{\hat{\mathbf{h}},\hat{\boldsymbol{\Omega}}} = \underset{\mathbf{f}}{\operatorname{argmax}}\, \mathrm{p}(\mathbf{g}|\hat{\boldsymbol{\Omega}}, \mathbf{f}, \hat{\mathbf{h}})\, \mathrm{p}(\mathbf{f}|\hat{\boldsymbol{\Omega}}). \tag{5.9}$$

We can also marginalise $\mathbf{h}$ and $\boldsymbol{\Omega}$ to obtain

$$\hat{\mathbf{f}} = \underset{\mathbf{f}}{\operatorname{argmax}} \int_{\mathbf{h},\boldsymbol{\Omega}} \mathrm{p}(\mathbf{g}|\boldsymbol{\Omega}, \mathbf{f}, \mathbf{h})\, \mathrm{p}(\mathbf{f}, \mathbf{h}|\boldsymbol{\Omega})\, \mathrm{p}(\boldsymbol{\Omega}) \mathrm{d}\mathbf{h} \cdot \mathrm{d}\boldsymbol{\Omega} \tag{5.10}$$

The two above inference models, are named Evidence- ((5.8) and (5.9)) and Empirical-based ((5.10)) analysis [154], respectively. The marginalised variables are also called hidden variables.

Note that the Evidence-based analysis can also be used to marginalise the image $\mathbf{f}$ as well as the

unknown parameters $\boldsymbol{\Omega}$ to obtain $p(\mathbf{h}|\mathbf{g})$, as in Equation (4.5), then the mode of this posterior distribution can be calculated. In order to estimate the original image $\mathbf{f}$ we can then use only the observation model. This is discussed further in Chapter 7 (see also [24]).

It is rarely possible to calculate in closed form the integrals involved in the Evidence- and Empirical-based Bayesian inference. To solve this problem we can use approximations of the integrands. Let us consider the integral in Equation (5.8), then for each value of $\mathbf{h}$ and $\boldsymbol{\Omega}$ we can calculate

$$\hat{\mathbf{f}}\Big|_{\mathbf{h},\boldsymbol{\Omega}} = \underset{\mathbf{f}}{\operatorname{argmax}}\, p(\mathbf{g}|\mathbf{f},\mathbf{h},\boldsymbol{\Omega})\, p(\mathbf{f}|\boldsymbol{\Omega}) \tag{5.11}$$

and perform the second-order Taylor's expansion of $\log p(\mathbf{g}|\mathbf{f},\mathbf{h},\boldsymbol{\Omega})$ around $\hat{\mathbf{f}}$. As a consequence of the approximation, the integral in Equation (5.8) is performed over a distribution on $\mathbf{f}$ that is Gaussian and usually easy to calculate. This methodology is called Laplace approximation [104, 133] and has been applied, for instance, by Galatsanos *et al.* [71, 72] to partially known blur deconvolution problems.

5.2.4 Expectation Maximisation

The expectation maximization (EM) algorithm, first described in [60] is an incredibly popular technique in signal processing for iteratively solving ML and MAP problems that can be regarded as having *hidden data*. Its properties are well studied: convergence to a *local* maximum of the likelihood or the posterior distribution is guaranteed. It is particularly suited to inverse problems in image restoration and BID as the unobserved image, $\mathbf{f}$, represents a natural choice for the hidden data and in consequence for solving Equation (5.8).

The EM algorithm has been used in BID, for example, in [108, 128] in a general frequency domain formulation and in [123] using the ARMA model (§4.4.2).

In its general form, as described in [60], the EM algorithm considers iteratively solving an associated problem using carefully chosen "complete" data, $\mathcal{X} = \{\mathcal{Y}, \mathcal{Z}\}$, composed of the observed data in the original problem termed the "incomplete" data, $\mathcal{Y}$, and the hidden variables $\mathcal{Z}$. The incomplete data is related to the complete data through a non-invertible contractive mapping, e.g. $\mathcal{Y} = [\mathbf{1},\mathbf{0}]\mathcal{X}$. That is to say, the (log) *marginal* likelihood $\mathcal{L}_{\mathcal{Y}}(\boldsymbol{\theta}) = \log p(\mathcal{Y};\boldsymbol{\theta})$

of the observed data given the parameters θ is found by marginalising the hidden variables:

$$p(\mathcal{Y};\,\theta) = \int_{\Omega_{\mathcal{X}|\mathcal{Y}}} p(\mathcal{X}\,|\,\theta)\;d\mathcal{X} = \int_{\Omega_{\mathcal{Z}}} p(\mathcal{Y}\,|\,\mathcal{Z};\theta)\,p(\mathcal{Z})\;d\mathcal{Z} \tag{5.12}$$

where the integration is over the domain of $\mathcal{X}$ that maps to $\mathcal{Y}$. The essential assumption is that given the complete data, the ML estimate of the parameters θ, found as $\arg\max_\theta \mathcal{L}_{\mathcal{X}}(\theta)$, is a comparatively easily solved problem. Here the (log) full likelihood is $\mathcal{L}_{\mathcal{X}}(\theta) = \log p(\mathcal{X}\,|\,\theta)$.

The problem is reduced to finding the *expected* ML estimate of θ given $\mathcal{X}$, but only actually given knowledge of $\mathcal{Y}$ and a previous estimate of the parameters. In practice, this estimate is found based upon the expected value of $\mathcal{L}_{\mathcal{X}}(\theta)$, or *sufficient statistics* that represent this distribution. This procedure is split into two steps, the Expectation, or E-step, and Maximisation or M-step, which are iterated to converge to the optimal ML solution for θ. Expressed mathematically, these steps are respectively:

$$\text{E Step:}\qquad \mathcal{L}(\theta;\hat{\theta}^{(k)}) = E\left[\log p(\mathcal{X}\,|\,\theta)\,\middle|\,\mathcal{Y};\hat{\theta}^{(k)}\right] = \int_{\Omega_{\mathcal{Z}}} \log p(\mathcal{X}\,|\,\theta)\,p\left(\mathcal{Z}\,\middle|\,\mathcal{Y};\hat{\theta}^{(k)}\right) d\mathcal{Z}, \quad \text{and} \tag{5.13}$$

$$\text{M Step:}\qquad \hat{\theta}^{(k+1)} = \arg\max_{\theta} \mathcal{L}\left(\theta;\hat{\theta}^{(k)}\right), \tag{5.14}$$

where $\hat{\theta}^{(k)}$ is the estimate of the parameters at the k^{th} iteration. If the function $\mathcal{L}(\theta;\hat{\theta}^{(k)})$ is a continuous function of θ and $\hat{\theta}$ then the algorithm is guaranteed to converge to a stationary point of the likelihood, $\mathcal{L}_{\mathcal{Y}}(\theta)$, though not necessarily the global optimum and hence the algorithm is sensitive to choice of initial conditions. At each step, the previous estimate of the parameters is effectively used to estimate some statistics of the complete data, which then gives an approximate version of the likelihood $\mathcal{L}_{\mathcal{X}}(\theta)$ in the E step that can be used to provide a new estimate of the parameters in the M step.

In the case of BID, when assuming a h is one of the parameters to be estimated in θ, the obvious choice for the complete data is $\mathcal{X} = \{\mathbf{g}, \mathbf{f}\}$. Given this choice, the parameter estimation in the M step becomes straightforward. Consider the case of a Gaussian posterior. In the E step, the resulting algorithm effectively estimates a regularised restored image $\hat{\mathbf{f}} = E[\mathbf{f}|\mathbf{g}, \hat{\theta}^{k}]$ given the previous parameter estimates, along with the covariance of this estimate, $\hat{\Sigma}_{\mathbf{f}|\mathbf{g}} = E\left[(\mathbf{f}-\hat{\mathbf{f}})^2\,\middle|\,\mathbf{g}, \hat{\theta}^{k}\right]$. In the M step, the resulting optimisation effectively uses these to calculate new estimates for the model parameters, $\theta = \{\mathbf{h}, \boldsymbol{\Omega}\}$. In the case of the ARMA model $\theta =$

$\{\mathbf{h}, \mathbf{a}, Q_v, Q_w\}$; this choice will be discussed further in §7.1.1.

5.2.4.1 Generalisations of EM

Some extensions and more general views of EM will now be considered. Firstly, note that the EM algorithm can be applied to MAP estimation as well as ML optimisation, simply by including the prior term on θ, and maximising $\mathcal{L}\left(\theta; \hat{\theta}^{(k)}\right) + \log p(\theta)$ in the M step.

A generalised expectation maximization (GEM) algorithm is a variant proposed in [60] whereby a partial M step is performed, i.e. a new estimate $\hat{\theta}^{(k+1)}$ is produced that does not maximise the current $\mathcal{L}(\theta; \hat{\theta}^{(k)})$, but simply increases its value, and consequentially also increases the marginal likelihood. Algorithms performing similar partial E steps were first discussed in [161].

An alternative view of the EM algorithm presented in [14, 161] is that it can be seen as a method for placing a *lower bound* on the marginal likelihood. It does this through the introduction of a bounding functional[1] $\mathcal{F}(q(\mathcal{Z}), \theta)$, and an auxiliary arbitrary distribution of the hidden variables, $q(\mathcal{Z})$. We can show the functional is a bound in that is always less than or equal to $\mathcal{L}_{\mathcal{Y}}(\theta)$, for any choice of $q(\mathcal{Z})$ [161]. It is defined as in (5.15):

$$\begin{aligned}
\mathcal{F}(q(\mathcal{Z}), \theta) &\triangleq \mathbb{E}\left[\log p(\mathcal{Y}, \mathcal{Z} \mid \theta)\right]_{q(\mathcal{Z})} - \mathbb{E}\left[\log q(\mathcal{Z})\right]_{q(\mathcal{Z})} && (5.15)\\
&= \mathbb{E}\left[\log p(\mathcal{Y} \mid \theta)\right]_{q(\mathcal{Z})} - \mathbb{E}\left[\log q(\mathcal{Z}) - \log p(\mathcal{Z} \mid \mathcal{Y}; \theta)\right]_{q(\mathcal{Z})} && (5.16)\\
&\equiv \mathcal{L}_{\mathcal{Y}}(\theta) - C_{KL}\big(q(\mathcal{Z}) \parallel p(\mathcal{Z} \mid \mathcal{Y}; \theta)\big) && (5.17)
\end{aligned}$$

where the Kullback-Leibler (KL) divergence $C_{KL}(\cdot \parallel \cdot)$, or relative entropy, between two distributions is by definition a non-negative quantity (this is shown via the Jensen inequality), with equality to zero only when both distributions are equal. Consequently, we have that $\mathcal{F}(q(\mathcal{Z}), \theta) \leqslant \mathcal{L}_{\mathcal{Y}}(\theta)$, for *any* distribution $q(\mathcal{Z})$, and the bound is tight (equal to the likelihood) when the function $q(\mathcal{Z}) = p(\mathcal{Z} \mid \mathcal{Y}; \theta)$. In this case, the first term of $\mathcal{F}(q(\mathcal{Z}), \hat{\theta})$ in (5.15) is equal to $\mathcal{L}(\theta; \hat{\theta})$.

With this interpretation, the E step can be seen as maximisation of the functional $\mathcal{F}(q(\mathcal{Z}), \hat{\theta}^{(k)})$ w.r.t the distribution $q(\mathcal{Z})$: i.e. at iteration $k+1$, setting $q^{k+1}(\mathcal{Z}) = p\left(\mathcal{Z} \middle| \mathcal{Y}; \hat{\theta}^{(k)}\right)$. The M step then maximises $\mathcal{F}(q^{k+1}(\mathcal{Z}), \theta)$ w.r.t θ (this is equivalent to (5.14), since the second term

[1] A functional is a function of a function — it is a mapping from a space of functions to a real number.

in (5.15) does not depend on θ). This reformulation of the EM process is a useful link to the VB methods described in the next section.

5.2.5 Variational Bayesian Approaches

The methods described so far, including the EM algorithm, have all calculated MAP or ML point estimates. Variational Bayesian (VB) methods[2] [14, 201] are a fairly recent addition to the family of Bayesian inference methods; they may be seen as generalisations of the EM algorithm. The EM algorithm has proven to be very useful in a wide range of applications; however, in many problems its application is not possible because the posterior distribution of the hidden variables $p(\mathcal{Z}|\mathcal{Y};\theta)$ is intractible, or cannot be specified in a standard form. Usually this is because its normalising constant, i.e. the marginal likelihood $p(\mathcal{Y}|\theta)$, cannot be evaluated. This posterior — or more precisely certain of its expectations, or expected sufficient statistics — are needed in order to perform the optimisation in the M step.[3] If none of these can be found then the standard EM algorithm cannot be applied. This becomes a problem for instance if we choose the hidden data as $\mathcal{Z} = \{\mathbf{f}, \mathbf{h}\}$, because the resulting posterior is no longer Gaussian in $\mathcal{Z}$, even if the individual priors are. In these cases, the VB approach can still be applied, by making use of an approximate posterior, $q(\mathcal{Z})$, in the place of the true one, and minimising the KL distance between the two as far as possible.

The second major advantage of VB methods is that they bring the EM framework fully in line with the Bayesian philosophy. That is, even if EM is applied using a MAP extension rather than to solve a ML problem, only point estimates of some of the parameters, θ, are found. Using VB, all parameters are now regarded as stochastic quantities, and their distributions are found, rather than just the distributions of the hidden parameters as with EM. Thus we may also examine the quality of the parameter estimates and not just their values. Note that the Laplace approximation [71, 72] can be considered as an intermediate step between inference based on the true posterior and the one based on a variational distribution approximation.

We will consider here the general case. In Chapter 9, the model used in [153] is exemplified before considering the application to the new extended models that will be developed. A quick

[2] These should not be confused with total variation based image models discussed in §B.4. Both are named due to results using variational calculus, but have entirely different intentions (in [8], both are used together for image restoration).

[3] Note that the normalising constant itself isn't actually required if the posterior is in a standard form such that the required statistics of the distribution can be evaluated for use in the M step.

point on the notation should be clarified. The full set of (stochastic) unknowns will now be denoted as $\boldsymbol{\Theta}$, with a particular unknown as θ, and the full set omitting variable θ as $\boldsymbol{\Theta}_{\setminus\theta}$. We can then write the approximate posterior as $q(\boldsymbol{\Theta})$. Some authors, (e.g. in [14, 132]) choose to still distinguish between the hidden variables and the parameters and retain "variational E" and "variational M" steps just as in the standard EM algorithm. When we are working with a hierarchical Bayesian model, drawing this distinction makes less sense; as such we will treat all unknown parameters equally as hidden variables with their own distribution. All these distributions are updated using the same type of "variational E" step.[4]

5.2.5.1 General Variational Bayesian Algorithm

We wish to perform inference using the posterior distribution, $p(\boldsymbol{\Theta} \mid \mathbf{g}) \propto p(g \mid \boldsymbol{\Theta})\, p(\boldsymbol{\Theta})$. However for the models under consideration, there is no analytic solution to maximise or find expectations of this distribution. Following the approach in [153], a separable analytic approximation of the posterior may be used:

$$p(\boldsymbol{\Theta} \mid \mathbf{g}) \approx q(\boldsymbol{\Theta}) = \prod_{\theta \in \boldsymbol{\Theta}} q(\theta). \tag{5.18}$$

This is sometimes termed the *mean field* approximation [14]. An iterative scheme may then be used to estimate the optimum distributions $q(\theta)$ for each variable θ.

The aim is to make the approximation as good as possible by minimising the KL divergence between the approximate distribution and the true posterior:

$$C_{KL}\big(q(\boldsymbol{\Theta}) \parallel p(\boldsymbol{\Theta} \mid \mathbf{g})\big) = \int_{\boldsymbol{\Theta}} q(\boldsymbol{\Theta}) \log\left(\frac{q(\boldsymbol{\Theta})}{p(\boldsymbol{\Theta} \mid \mathbf{g})}\right) d\boldsymbol{\Theta}. \tag{5.19}$$

As already mentioned, $C_{KL}(\cdot \parallel \cdot)$ is always nonnegative, and equal to zero only for $q(\boldsymbol{\Theta}) = p(\boldsymbol{\Theta} \mid \mathbf{g})$, which is the EM result. In this case we will substitute the separable approximation.

The KL divergence will now be optimised with respect to a particular variable of interest, θ. To begin, Bayes' rule is applied to the posterior, and the term $p(\mathbf{g})$ which does not depend on

[4] This may simplify however, if they are at the top of our tree-structured graphical model; this is the only difference described in [14] in the "variational M" step

θ may be regarded as a constant:

$$C_{KL}\big(q(\boldsymbol{\Theta}) \parallel p(\boldsymbol{\Theta} \mid \mathbf{g})\big) = \int_{\boldsymbol{\Theta}_{\setminus\theta}} \int_{\theta} q(\boldsymbol{\Theta}) \log\left(\frac{q(\boldsymbol{\Theta})}{p(\mathbf{g} \mid \boldsymbol{\Theta})\, p(\boldsymbol{\Theta})}\right) d\theta d\boldsymbol{\Theta}_{\setminus\theta} + \text{const.} \tag{5.20}$$

Now, the separable approximation (5.18) may be applied, and those $q(\,)$ distributions which do not depend on θ may be taken outside the inner integral:

$$\begin{aligned} C_{KL}\big(q(\boldsymbol{\Theta}) \parallel p(\boldsymbol{\Theta} \mid \mathbf{g})\big) &= \int_{\boldsymbol{\Theta}_{\setminus\theta}} q(\boldsymbol{\Theta}_{\setminus\theta}) \int_{\theta} q(\theta) \log\left(\frac{q(\boldsymbol{\Theta}_{\setminus\theta})\, q(\theta)}{p(\mathbf{g} \mid \boldsymbol{\Theta})\, p(\boldsymbol{\Theta})}\right) d\theta\, d\boldsymbol{\Theta}_{\setminus\theta} + \text{const.} \\ &= \int_{\boldsymbol{\Theta}_{\setminus\theta}} q(\boldsymbol{\Theta}_{\setminus\theta}) \left(\log q(\boldsymbol{\Theta}_{\setminus\theta}) \int_{\theta} q(\theta)\, d\theta \right. \\ &\quad \left. + \int_{\theta} q(\theta) \log\left(\frac{q(\theta)}{p(\mathbf{g} \mid \boldsymbol{\Theta})\, p(\boldsymbol{\Theta})}\right) d\theta \right) d\boldsymbol{\Theta}_{\setminus\theta} + \text{const.} \end{aligned} \tag{5.21}$$

$$\begin{aligned} &= \overbrace{\int_{\boldsymbol{\Theta}_{\setminus\theta}} q(\boldsymbol{\Theta}_{\setminus\theta}) \log q(\boldsymbol{\Theta}_{\setminus\theta}) d\boldsymbol{\Theta}_{\setminus\theta}}^{\text{independent of } q(\theta)} \overbrace{\int_{\theta} q(\theta)\, d\theta}^{=1} + \text{const.} \\ &\quad + \int_{\boldsymbol{\Theta}_{\setminus\theta}} q(\boldsymbol{\Theta}_{\setminus\theta}) \int_{\theta} q(\theta) \log\left(\frac{q(\theta)}{p(\mathbf{g} \mid \boldsymbol{\Theta})\, p(\boldsymbol{\Theta})}\right) d\theta\, d\boldsymbol{\Theta}_{\setminus\theta}, \end{aligned} \tag{5.22}$$

such that only the final term depends on $q(\theta)$, and the other terms may be regarded as constants for the purposes of minimisation. To obtain the optimal distribution $q(\boldsymbol{\Theta})$, we find for each θ:

$$\widehat{q}(\theta) = \underset{q(\theta)}{\operatorname{argmin}}\, C_{KL}\big(q(\boldsymbol{\Theta}) \parallel p(\boldsymbol{\Theta} \mid \mathbf{g})\big) \tag{5.23}$$

$$= \underset{q(\theta)}{\operatorname{argmin}}\, \mathbb{E}\left[\int_{\theta} q(\theta) \log\left(\frac{q(\theta)}{p(\mathbf{g} \mid \boldsymbol{\Theta})\, p(\boldsymbol{\Theta})}\right) d\theta\right]_{q(\boldsymbol{\Theta}_{\setminus\theta})}. \tag{5.24}$$

Differentiating the argument with respect to $q(\theta)$ gives the result:

$$\boxed{\log \widehat{q}(\theta) = \text{const.} + \mathbb{E}\left[\log\big(p(\mathbf{g} \mid \boldsymbol{\Theta})\, p(\boldsymbol{\Theta})\big)\right]_{q(\boldsymbol{\Theta}_{\setminus\theta})}} \tag{5.25}$$

The constant appears under the constraint $\int q(\theta)\, d\theta = 1$. The differentiation in (5.24) is actually carried out by variational calculus (functional differentiation) — this explains the name of the VB approach. Further details are given in [14, 144], but a simpler way to view

the optimisation is to consider rewriting (5.24) as

$$\widehat{q}(\theta) = \underset{q(\theta)}{\operatorname{argmin}} \left(\int_{\theta} q(\theta) \left(\log q(\theta) - \mathbb{E}\left[\log \left(p(\mathbf{g} \mid \boldsymbol{\Theta}) p(\boldsymbol{\Theta}) \right) \right]_{q(\boldsymbol{\Theta}_{\setminus\theta})} \right) d\theta \right). \tag{5.26}$$

It should then be clear that (5.25) minimises the expression, as it is in fact the KL divergence between $q(\theta)$ and $\exp\left(\mathbb{E}\left[\log \left(p(\mathbf{g} \mid \boldsymbol{\Theta}) p(\boldsymbol{\Theta}) \right) \right]_{q(\boldsymbol{\Theta}_{\setminus\theta})} \right)$.

Equation (5.25) tells us how to derive the optimal distribution for $\widehat{q}(\theta)$ given knowledge of the other distributions. By iterating through each θ one at a time and solving (5.25) for the new distribution $\widehat{q}(\theta)$, we have an iterative procedure whereby the KL divergence is decreased optimally at each step; this is termed the iterative variational Bayesian (IVB) procedure in [201]. In the IVB procedure we will indicate the distribution $\widehat{q}(\theta)$ at a particular iteration, k, by $q^k(\theta)$. Finally it should be noted, however, that the IVB procedure may still converge to a local optima in the space of approximate distributions, just as with the EM algorithm.

5.2.5.2 Discussion

The approach just described is a *free-form* factorisation: the variational optimisation of the KL divergence dictates the functional form of the resulting distributions. We may also choose a sub-optimal improvement at each iteration, so long as the KL divergence is decreased at each step. This is in analogy to a GEM algorithm. For example, a particular functional form such as a Gaussian with a known variance may be prescribed, then the optimisation for this $q(\cdot)$ distribution is reduced to finding its mean.

Alternating minimization strategies can also be employed. For example, in [132], a cascaded EM algorithm is proposed similar to the AM algorithm, where at each iteration the distributions of the parameters $\mathbf{f}$ and $\mathbf{h}$ are calculated in an alternating fashion while assuming one parameter to be constant. This approach can also be interpreted as an EM algorithm where at each stage the current estimate of one parameter is assumed "known" in estimating the other parameter, as in the classical image restoration problems.

Just as with the EM algorithm, another interpretation of the whole procedure, discussed in [14, 132], is that a lower bound $\mathcal{F}(q(\theta_1), q(\theta_2), \cdots)$ of the marginal likelihood is being provided by the approximate distributions, and the bound is being optimised through the iterative procedure. The difference is that now the bound is optimised with respect to the

distributions of the parameters, and the bound is never made tight as in the E step of the EM algorithm, but simply increased each time a $q(\cdot)$ distribution is updated.

As discussed in [14], the use of *conjugate exponential* models, that is distributions from the exponential family along with conjugate priors, greatly facilitates the application of VB in a standardised fashion; in fact as will be seen in Chapter 9, it is just the sufficient statistics, or moments, of these distributions that need be calculated at each step. The particular advantage is that a free-form update of the approximate distributions will always result in a new $q(\theta)$ distribution that is still a member of the exponential family. Other distributions may result in more unwieldy results where the approximate form changes at each iteration.

5.2.5.3 Application to BID problem

The first application of VB to the BID problem was in a slightly different context by Miskin and MacKay [143], where the problem was addressed as an extension to a blind source separation (BSS) problem — essentially a multichannel observation with a mixing between the source images. This method used a sparsity inducing Laplace prior directly on the images and blurs; this is not suitable for natural images because it does not model the correlations in the image. Later, Likas *et al.*[132] and Molina *et al.*[153] have used Gaussian image and blur priors as described in §4.4.1 which do model these correlations. In [153], the hierarchical Bayesian framework is also used to enable hyperparameter estimation (see Chapter 9 for further details). Recently, Tzikas *et al.*[217] used a similar hierarchical formulation with a kernel based PSF model and a nonstationary image prior, like the one to be described in §9.4. Fergus *et al.*[67] have built upon the work by Miskin and MacKay, by applying an empirically trained sparse (heavy-tailed) prior over the image gradients and the blur, which is similar to using edge-preserving nonstationary models.

The expression for distribution updates (5.25) is actually fairly intuitive once it is applied to a particular hierarchical Bayesian network, as will be seen. In general, observing that the term inside the expectation is the log of the global joint distribution, this may be factorised into the sum of the log of all the conditional distributions appearing in the tree. Then for a given variable θ_i, only those conditional terms that include θ_i need be considered (as the other terms will just form part of the normalising constant for $q(\theta_i)$). These terms to be included are the conditional distribution of the child variable to θ_i in the tree, and the distribution of θ_i conditional on its parents (i.e. its prior). Furthermore only those terms which depend on

other unknowns ($\boldsymbol{\Theta}_{\setminus \theta_i}$) remain inside the expectaiton; hence if θ_i is a hyperparameter at the top of the tree, its hyperprior may be taken outside the expectation. The examples in Chapter 9 will make these concepts more clear.

5.2.6 Sampling Methods: Markov chain Monte Carlo

The most general approach to performing inference for the BID problem is to simulate the posterior distribution in Equation (4.1). This in theory allows us to perform inference on arbitrarily complex models in high-dimensional spaces, where no analytic solution is available. Markov chain Monte Carlo (MCMC) methods (see, e.g., [6, 78, 160, 167]) attempt to approximate the posterior distribution by the statistics of samples generated from a Markov chain.

MCMC methods have been targeted at such problems in Bayesian inference, where typically one needs to evaluate integrals, or expectations, in high dimensional spaces. For some unknown vector in a high dimensional space, $\mathbf{x} \in \mathcal{X}$, given a known $\mathbf{y}$, we need to find

$$\Phi = \mathbb{E}\left[\phi(\mathbf{x})\right]_{p(\mathbf{x}|\mathbf{y})} = \int_{\mathcal{X}} \phi(\mathbf{x})\, p\left(\mathbf{x}\,|\,\mathbf{y}\right)\, d\mathbf{x}$$

For instance, one parameter estimate we are interested in is the MMSE estimate, $\mathbb{E}\left[\mathbf{x}\right]_{p(\mathbf{x}|\mathbf{y})}$. Another situation where high dimensional integrals have to be evaluated is if we need to know the normalisation constant (partition function) in the posterior distribution, for instance in model comparison. As it turns out however, we do not need to know this constant for the purposes we are interested in.

5.2.6.1 Monte Carlo Integration

One numerical method to perform integrals of arbitrary functions is to discretise the space $\mathcal{X}$ and evaluate the function by a summation of the function values over each cell. However this soon becomes impractical in higher dimensions, as the number of cells grows exponentially with dimension.

Monte Carlo methods propose a stochastic approach to these types of integration problem. Assuming we have a set of samples from the distribution available, $\{\mathbf{x}^{(r)}\}_{r=1}^{R} \leftarrow p\left(\mathbf{x}\,|\,\mathbf{y}\right)$, we

can approximate the integral by a simple average of the function value at these samples:

$$\hat{\Phi} = \frac{1}{R} \sum_{r} \phi(\mathbf{x}^{(r)}).$$

In the limit as $\mathbb{R} \to \infty$, by the law of large numbers, we find that $\hat{\Phi} \to \mathbb{E}\left[\phi(\mathbf{x})\right]_{p(\mathbf{x}|\mathbf{y})}$. More importantly, we find that if $\sigma_\phi^2 \triangleq \mathrm{Var}(\phi(\mathbf{x})) = \mathbb{E}\left[\phi(\mathbf{x} - \Phi)^2\right]_{p(\mathbf{x}|\mathbf{y})}$, then $\mathrm{Var}(\hat{\Phi}) = \frac{\sigma_\phi^2}{R}$, i.e. the variance of the estimator is independent of dimension, but only depends on the number of samples drawn. This is an important results in order to evaluate the high dimensional integrals that we are interested in.

5.2.6.2 Markov chains

In order to be able to perform Monte Carlo integration, we need to be able to draw the samples. This turns out to be surprisingly difficult for most non-elementary distributions. In 1D we can use probability transformations of uniform variables to generate for instance scalar Gaussian or inverse Gamma distributed variables. Other 1D distributions can be also found using a look-up table with the inverse-CDF method to transform uniform random variables. In low dimensions we can use techniques such as rejection sampling to approximately bound the distribution we wish to sample from by another from which we can. Unfortunately this method also becomes very inefficient in higher dimensions, as whenever the distributions are not a good match we end up rejecting a lot of samples.

It turns out that we can instead, however construct *Markov chains* to draw samples from the desired distribution. A Markov chain is a discrete-time stochastic process, such that $p\left(\mathbf{x}^{(i)} \,\middle|\, \mathbf{x}^{(i-1)}, \cdots, \mathbf{x}^{(1)}\right) = p\left(\mathbf{x}^{(i)} \,\middle|\, \mathbf{x}^{(i-1)}\right)$, i.e. samples at a time i only depend on the value of the state of the sample at the previous time step. We define the probability of being in the new state $\mathbf{x}_c$ given a previous state $\mathbf{x}_r$, $p\left(\mathbf{x}^{(i)} = \mathbf{x}_c \,\middle|\, \mathbf{x}^{(i-1)} = \mathbf{x}_r\right)$, as the transition probability. All of these possible transitions define a *transition kernel* $T\left(\mathbf{x}^{(i)} \,\middle|\, \mathbf{x}^{(i-1)}\right) = p\left(\mathbf{x}^{(i)} \,\middle|\, \mathbf{x}^{(i-1)}\right)$.

Let us consider for simplicity first the example of a finite state space. Then the kernel becomes a matrix $\mathbf{T}$, where the transition probability just defined is located at row and col (r, c) of the matrix. This matrix has the stochastic property, that all rows sum to one (to represent a certainty of transitioning to some state at each time step). The entries in this matrix represent a state transition graph showing the chance of moving from one state to another. For example

the matrix

$$\mathsf{T} = \begin{bmatrix} 0.1 & 0.9 & 0 \\ 0.5 & 0.2 & 0.3 \\ 0.1 & 0.7 & 0.2 \end{bmatrix} \equiv \mathsf{T}\left(x^{(i)} \,\middle|\, x^{(i-1)}\right)$$

gives the state transition graph shown in fig. B.1.

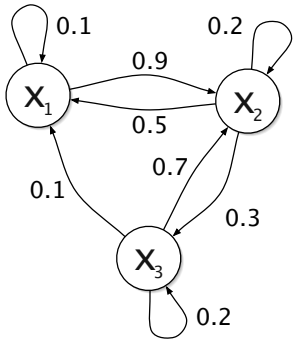


Figure 5.1: *Example of a state transition diagram for a discrete state space Markov chain*

Now we wish to consider Markov chains where a couple of properties of the graph hold. *Aperiodicity* specifies that repeated application of the transition matrix does not create periodic cycles through the states in the graph. *Irreducibility* requires that all states are accessible from all others; i.e. there is a positive probability of visiting each state finitely often. Assuming these properties, which are easily satisfied in practice, it can be shown that the Markov chain will converge to a *stationary* or *invariant* distribution.

It turns out that in the finite case, this corresponds to having all eigenvectors having magnitudes between zero and one, with exactly one taking the value 1. Given an current state, in order to find the new state, we just marginalise: $p\left(\mathbf{x}^{(i)}\right) = \sum_{\mathbf{x}^{(i-1)}} p\left(\mathbf{x}^{(i-1)}\right) T\left(\mathbf{x}^{(i)} \,\middle|\, \mathbf{x}^{(i-1)}\right)$. In the finite state case, this is just matrix multiplication,

$$\boldsymbol{\pi}(i) = \boldsymbol{\pi}(i-1)\mathbf{T},$$

where $\boldsymbol{\pi}(i) = [p\left(\mathbf{x}^{(i)}\right)]$, the row vector of probabilities of being in each state, and typically we start off in a known state, e.g. $\boldsymbol{\pi}(1) = [1, 0, 0]$. Based upon the eigenvalues, we find that repeated application of the transition matrix, $\boldsymbol{\pi}(i+n) = \boldsymbol{\pi}(i)\mathbf{T}^n$, results in the invariant distribution $\boldsymbol{\pi}^*$, such that $\boldsymbol{\pi}^* = \boldsymbol{\pi}^*\mathbf{T}$, i.e. $\boldsymbol{\pi}^*$ is the left eigenvector of the kernel with eigenvalue one. This is the important property of Markov chains which enables us to use them to draw samples from the distribution $\boldsymbol{\pi}^*$.

Of course we are interested in dealing with continuous valued variables, which correspond to infinite dimensional Markov chains. However the notions described above still hold, with the matrix replaced by a kernel function, and the stationary vectors being the stationary distribution. The sumations are replaced by integrals, for instance the state transition equation, also termed the Chapman-Kolomogrov equation, becomes

$$p\left(\mathbf{x}^{(i)}\right) = \int p\left(\mathbf{x}^{(i-1)}\right) T\left(\mathbf{x}^{(i)} \middle| \mathbf{x}^{(i-1)}\right) d\mathbf{x}^{(i-1)}. \tag{5.27}$$

5.2.6.3 Markov chain Monte Carlo

Based upon the previous ideas, if we want to sample a distribution $\pi^*(\mathbf{x})$, all we have to do is find a valid transition kernel which yields this distribution as its stationary distribution. Fortunately, this is easier than it may sound, as methods have been designed to construct the kernels, or more straightforwardly, how to apply them, to yield the correct Markov chain. In its limit this Markov chain will give independent samples from the required distribution, which can then be used for Monte Carlo integration to evaluate the required integrals or expectations. In practice, this means that the samplers must be run for a certain *burn-in* period before the samples become from the correct distribution.

The first method proposed was the *Metropolis* method [140], which is generally applicable to drawing samples from any PDF, so long as we can evaluate it up to a constant of proportionality. A *proposal* distribution is defined $q\left(\mathbf{x}^{i} \middle| \mathbf{x}^{i-1}\right)$, that we know how to sample from. Typically this distribution depends on the most recent value of the sample, $\mathbf{x}^{i-1}$. In the Metropolis method the kernel must be symmetric, i.e. $q\left(\mathbf{x}^{(i-1)} \middle| \mathbf{x}^{(i)}\right) = q\left(\mathbf{x}^{(i)} \middle| \mathbf{x}^{(i-1)}\right)$. An extension, the *Metropolis-Hastings* method [91] allows for non-symmetric kernels. The principle of either method is to accept or reject the new sample, with a certain probability α, which depends on the ratio of the old and proposed values of the functions $q(\cdot|\cdot)$ and $\pi^*(\cdot)$. To ensure efficiency of the method, we need to make the acceptance probability as high as possible. There is a trade off between moving quickly around the space versus having the samples rejected. Thus the task of designing a good proposal distribution is important, and problem dependant.

5.2.6.4 Gibbs Sampler

A special case of the Metropolis-Hastings method is the *Gibbs sampler* [76], which uses a kernel that changes at each step. This method has become particularly popular for applications in Bayesian inference, because of a couple of nice features. Firstly, the acceptance probability is always one, meaning samples aren't wasted. Secondly, apart from the choice of how to partition the variables, the kernel is automatically defined by the problem, and the resulting proposal distributions are just the full conditional distributions of each variable from the posterior,

$$q\left(\theta^{(i)} \,\middle|\, \boldsymbol{\Theta}^{(i-1)}\right) = p\left(\theta \,\middle|\, \boldsymbol{\Theta}_{\setminus\theta}{}^{(i-1)}\right).$$

If we can write down analytic expressions for the *conditional* distributions of all the parameters we wish to estimate, given the others, we simply draw samples from each of the distributions in turn, conditioned on the most recently sampled values of the other parameters. For example if we want to simulate $p(\mathbf{f}, \mathbf{h}, \boldsymbol{\Omega} \mid \mathbf{g})$, the iterations would proceed as follows:

$$\begin{aligned}
\text{First iteration:} \quad \mathbf{f}^{(1)} &\leftarrow p(\mathbf{f} \mid \mathbf{h}^{(0)}, \boldsymbol{\Omega}^{(0)}, \mathbf{g}) \\
\mathbf{h}^{(1)} &\leftarrow p(\mathbf{h} \mid \mathbf{f}^{(1)}, \boldsymbol{\Omega}^{(0)}, \mathbf{g}) \\
\boldsymbol{\Omega}^{(1)} &\leftarrow p(\boldsymbol{\Omega} \mid \mathbf{f}^{(1)}, \mathbf{h}^{(1)}, \mathbf{g}) \\
\text{Second iteration:} \quad \mathbf{f}^{(2)} &\leftarrow p(\mathbf{f} \mid \mathbf{h}^{(1)}, \boldsymbol{\Omega}^{(1)}, \mathbf{g}) \\
\mathbf{h}^{(2)} &\leftarrow p(\mathbf{h} \mid \mathbf{f}^{(2)}, \boldsymbol{\Omega}^{(1)}, \mathbf{g}) \\
\boldsymbol{\Omega}^{(2)} &\leftarrow p(\boldsymbol{\Omega} \mid \mathbf{f}^{(2)}, \mathbf{h}^{(2)}, \mathbf{g}) \\
&\vdots \qquad \vdots \\
t^{\text{th}} \text{ iteration:} \quad \mathbf{f}^{(t)} &\leftarrow p(\mathbf{f} \mid \mathbf{h}^{(t-1)}, \boldsymbol{\Omega}^{(t-1)}, \mathbf{g}) \\
&\vdots \qquad \vdots
\end{aligned}$$

where the symbol $\leftarrow$ means the value is drawn from the distribution on the right. Notice the similarity to the iterative procedure for the Variational Bayesian approach in Equation (5.24), where instead of drawing samples we are taking expectations of the same distributions. Similarly, with the use of *Simulated Annealing* [76], the ICM formulation can be considered a deterministic approximation of the sampling, where the conditional distributions are replaced by degenerate distributions at their modes (termed "instantaneous freezing" in [20]).

Once we have accumulated samples, point estimates and other statistics of the distribution may be found using Monte Carlo integration; for example, to find the MMSE estimate of the $\mathbf{f}$ we simply take the mean of the samples, $\frac{1}{n}\sum_{t=1}^{n}\mathbf{f}^{(t)}$.

Clearly, these methods can provide solutions closer to the optimal one than AM or any of the other methods. However, they are very computationally intensive by comparison, and although under certain mild assumptions convergence to the posterior is theoretically guaranteed, in practice it can be hard to tell when this has occurred; it may take a long time to explore the parameter space.

5.3 Conclusions

In this chapter, the various Bayesian inference methods that can be used to tackle the BID problem have been explored. BID methods appearing in the literature have been surveyed and categorised. By drawing equivalencies under the Bayesian inference methods, we were able to show the similarities and differences between a number of methods, even some not originally formulated this way. Details of how the different inference methods can be applied to BID were outlined, with some additional background on the variational Bayesian (VB) methodology and MCMC that will be applied in Chapters 9 and 10 given.

The next chapter begins the second part of the thesis, in which novel solutions to the BID problem are proposed. New models are developed in Chapter 6, which are then applied to the problem through different Bayesian inference methods in the remaining chapters.

Part II

Proposed Methodology

Chapter 6
Proposed Models

It has been seen in §4.4 that a large number of image models have been proposed in the past for the BID problem and other inverse problems in image restoration. Because the BID problem is even more ill-posed than regular image restoration, being able to reach a reasonable solution is really highly dependent on the quality of the prior information we can apply. More than just improving quality, accurate modelling of the true image and blur will also help BID algorithms to converge more quickly and reliably, as we restrict the space of possible solutions to ones that make sense. With poor priors, BID algorithms may even diverge away from the solution. So, by designing new priors it should be possible to improve restoration performance when the blur is unknown.

Many existing models in non-blind image restoration have utilised the idea of edge preservation; that is to vary the strength of the prior in smooth regions and near edges, to avoid over-smoothing details. It is only more recent works, e.g. [50, 57, 67, 217], that such concepts have been applied to BID problems, rather than using non-adaptive global smoothness priors. Nevertheless, these local models all operate on a very localised basis, and can only model the degree (and possibly direction) of the edge transitions. That is, if we view each of these models as a particular type of hierarchical GMRF, it is only the *first order* spatial interactions that are being utilised in each case. These models therefore cannot respond to more subtle texture and details in the image.

The models investigated here will use higher order interactions over larger support regions, allowing for texture and other features to be modelled as well as simply distinguishing edges which occur between immediately neighbouring pixels. As well as improving the restoration of these features in the image, using such a model should also help the solution of the BID problem. Compared to more simple models that would tend to smooth over the textured regions, or replace them with a simple step transition, having them available will aid blur estimation. This is because blur estimation typically involves a deconvolution step with the roles of image and blur inverted. The extra image content will provide information about

frequency components that isn't present in the over-smoothed version, aiding restoration of these frequencies in the PSF.

In this work, we build up a nonstationary AR image model using standard conjugate exponential distributions to provide a tractable (via the tools described in §5.2) analysis under the hierarchical Bayesian framework. Note that these models may also be classed as a type of MRF. The distinction between smooth, textured and edge regions has also been made in [233], where a cost function approach was used, and the locally adaptive regularisation parameters estimated based upon a segmentation into clusters of these three types of regions. However, this method also does not model the higher order correlations between pixels. In the proposed methods, rather than partitioning *a priori* into smooth, textured, or edge regions, we let the model adapt to the underlying structure of the image through the estimation of local texture features.

Part of the motivation for the models used in this work comes from similar models used for blind restoration of audio signals in [92, 93]. Firstly, one of the main themes explored in [92] was that nonstationary models can aid identifiability of model parameters when there is ambiguity, such as in the case of a convolutive mixture. Secondly, the use of block-stationary AR (BSAR) signal models has been examined, and these models are adapted here for image modelling. There are, obviously, significant differences between the types of signals encountered in audio — speech, music, and so on — and photographic images. The extension from the 1D case to 2D is one of the problems that is addressed in this work. The concept of *causality*, which is meaningful for time-based media, holds no natural equivalent for spatial processes such as images. Furthermore, images and PSFs are positive valued quantities, exhibiting local non-zero means, unlike audio signals. Hence the models have had to be adapted to these requirements.

Non-causal AR models were studied and successfully applied for use in texture synthesis [55, 103] and inpainting of missing image details [113]. These existing studies provide a further reason that they are suitable for representing texture and local features in natural image models. A previous attempt at using nonstationary AR models for image denoising was given in [32]. There are some resemblances to the models presented in this chapter, however only second-order processes were used, and these were parameterised by slowly varying basis functions rather than being estimated locally, and recursive estimation rather than direct estimation was used. Another way to view the models is from the point of synthetic data generation, and

misfit of the residual on real data: the smaller the residual error, the better the model can fit a real image; this is discussed in §6.2.

The AR model along with a general linear convolutive blur results in an ARMA formulation; this was used previously in the BID problem [107, 123, 210], although in a stationary sense, with just a few parameters estimated to model the whole image globally. Furthermore either a causal image model was assumed, or a circulant approximation employed across the whole image. We also build upon these ideas from these models and combine them with the advantages of using a nonstationary model to adapt locally to different image features.

The remaining chapters of the thesis follow a chronological development of these models and their inference using three different Bayesian methods. The inference methods themselves dictate some of the complexity that may be designed into the model, and so the two are inextricably tied together. As such we begin in Chapter 7 with the simplest form of the model, which is a simple 2D adaptation of the causal BSAR model used in [93], in which the observation noise is also ignored, facilitating tractable marginalisation. This method, although not equipped with a realistic observation model, provides some useful conceptual examples and shows the improvements in blur estimation that are possible from a nonstationary model. A more realistic observation model is then considered, including the problem of including missing data that appears at the image boundaries; this is necessary for application to real degraded images in Chapter 10. With the variational Bayesian (VB) or Markov chain Monte Carlo (MCMC) methods in Chapter 9 and Chapter 10, simultaneous estimation of the image is also possible, along with adaptive regularisation based upon the estimated BSAR model. The complete model is then built up, beginning from the stationary case, then adding in turn local variance estimation, texture modelling, means, and positivity.

As discussed in §4.3.4, new blur models have not been explicitly designed in this thesis. Instead, priors similar to the new image models are applied (with suitable differences in the hyperparameters, for instance to account for the different smoothness that is expected in a point-spread function (PSF)). This indicates the adaptability of the proposed models, and gives adequate results in practice.

After introducing the principles of the block-stationary AR (BSAR) model in §6.2, the mathematical formulation of the image model without local means is summarised in §6.3 and then reformulated with the means in §6.4. Following this the complete hierarchical model includ-

ing the hyperpriors is described.

6.1 Review of stationary ARMA model based and ML methods

The combination of a stationary AR image model followed by a convolution with a PSF can be interpreted as an ARMA model. With this notion, BID is reduced to an ARMA parameter estimation problem [123, 124, 210], for which several solutions have been proposed. We will use this as a starting point to describe the improvements that may be obtained by moving towards a nonstationary version of this model.

This model is illustrated diagrammatically as a generative process in fig. 6.1 where the AR and MA terms are represented as filters (the MA part is an all-zero filter, generally the terms MA and filter are used interchangeably). The figure shows a synthetic realisation of the image field, $\mathbf{f}$, being passed through the standard observation model. The AR parameters, $\boldsymbol{\alpha}_f$, and variances, σ_v^2, in each part of the image are constant. This generative model can also represent the simultaneous autoregression (SAR) model used in [153], if the parameters $\boldsymbol{\alpha}_f$ are fixed to represent the discrete Laplacian.

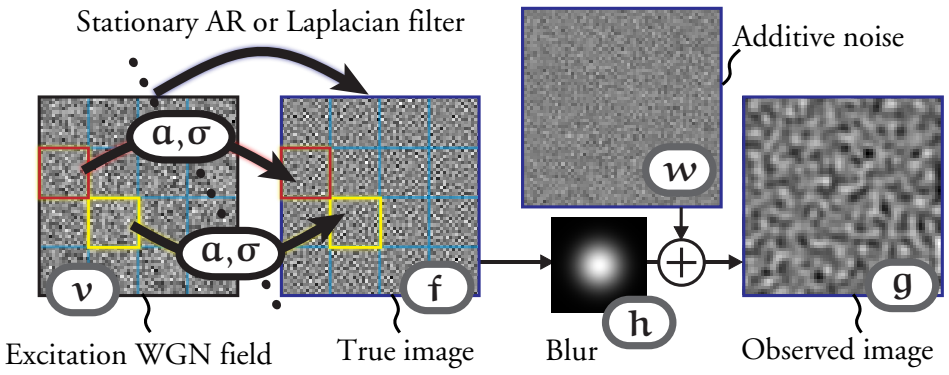


Figure 6.1: *Stationary ARMA model for the source and observed (blurred) images. Each region in the image is assumed to have the same AR model and blur model.*

Based upon the source and observation equations (4.35b) and (2.7) in matrix vector form, the PDFs of the source and observed image may be found. The prior PDF representing the true image is found from a linear transformation $\boldsymbol{v} = (\mathbf{I} - \mathbf{A})\mathbf{f}$ of the excitation, $\boldsymbol{v} \sim \mathcal{N}(0, \mathbf{Q}_v)$ (where $\mathbf{Q}_v = \sigma_v^2\mathbf{I}$ in this stationary case) , as

$$\mathrm{p}\left(\mathbf{f}\,\middle|\,\mathbf{a}_\mathrm{f}, \sigma_v^2\right) = \mathcal{N}(\mathbf{f}|0, (\mathbf{I}-\mathbf{A})^{-1}\mathbf{Q}_v(\mathbf{I}-\mathbf{A})^{-\mathsf{T}}) \tag{6.1}$$

$$= \sqrt{\frac{\det|\mathbf{I}-\mathbf{A}|^2}{(2\pi)^{\mathrm{L}_+}\det|\mathbf{Q}_v|}}\exp\left[-\frac{1}{2}\mathbf{f}^\mathsf{T}(\mathbf{I}-\mathbf{A})^\mathsf{T}\mathbf{Q}_v^{-1}(\mathbf{I}-\mathbf{A})\mathbf{f}\right]. \tag{6.2}$$

6.1.1 ML Parameter Estimation

The output of the ARMA model is found using the standard observation model, (2.7). Hence the likelihood of $\mathbf{g}$ (this is the marginal likelihood discussed in §5.2.4 and §7.1.1) may be found by using (4.3) and noting that addition of two independent random variables results in the convolution of their PDFs, or for zero-mean Gaussian variables, the addition of their covariances:

$$\mathrm{p}\left(\mathbf{g}\,\middle|\,\mathbf{a}_\mathrm{f}, \mathbf{h}, \sigma_v^2, \sigma_w^2\right) = \mathrm{p}\left(\mathbf{H}\mathbf{f}\right) * \mathrm{p}\left(\boldsymbol{w}\right) \tag{6.3}$$

$$= \mathcal{N}(\mathbf{g}\mid 0, \mathbf{H}(\mathbf{I}-\mathbf{A})^{-1}\mathbf{Q}_v(\mathbf{I}-\mathbf{A})^{-\mathsf{T}}\mathbf{H}^\mathsf{T}+\mathbf{Q}_w) \tag{6.4}$$

$$= \frac{1}{\sqrt{(2\pi)^{\mathrm{L}_\mathrm{g}}\det|\mathbf{H}\boldsymbol{\Sigma}_\mathrm{f}\mathbf{H}^\mathsf{T}+\mathbf{Q}_w|}}\exp\left[-\frac{1}{2}\mathbf{g}^\mathsf{T}(\mathbf{H}\boldsymbol{\Sigma}_\mathrm{f}\mathbf{H}^\mathsf{T}+\mathbf{Q}_w)^{-1}\mathbf{g}\right]. \tag{6.5}$$

The terms $\mathrm{E}[\mathbf{f}\mathbf{f}^\mathsf{T}] = (\mathbf{I}-\mathbf{A})^{-1}\mathbf{Q}_v(\mathbf{I}-\mathbf{A})^{-\mathsf{T}}$ and $\mathrm{E}[\mathbf{g}\mathbf{g}^\mathsf{T}] = \mathrm{H}(\mathbf{I}-\mathbf{A})^{-1}\mathbf{Q}_v(\mathbf{I}-\mathbf{A})^{-\mathsf{T}}\mathbf{H}^\mathsf{T}+\mathbf{Q}_w$ are denoted $\boldsymbol{\Sigma}_\mathrm{f}$ and $\boldsymbol{\Sigma}_\mathrm{g}$ respectively for brevity.

To find the maximum likelihood (ML) solution, it is usual to take logs and re-express (6.5) as the log-likelihood of $\boldsymbol{\theta}$ given $\mathbf{g}$, $\mathcal{L}_\mathbf{g}(\boldsymbol{\theta})$. The ML solution is found as the solution to the parameter vector, $\boldsymbol{\theta} \triangleq \{\mathbf{a}, \mathbf{h}, \boldsymbol{\sigma}_v, \sigma_w^2\}$, which maximises the log-likelihood expression:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}_\mathbf{g}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \left(-\ln(\det|\boldsymbol{\Sigma}_\mathrm{g}|) - \mathbf{g}^\mathsf{T}\boldsymbol{\Sigma}_\mathrm{g}\mathbf{g}\right) \tag{6.6}$$

This is a highly nonlinear function of the parameters and cannot be simplified due to the addition of the $\mathbf{Q}_w$ term. In order to proceed, it is either necessary to make some approximations or to use iterative or numerical methods. Lagendijk *et al.*[123, 124] have considered different methods for maximising the log-likelihood in [124], including Gradient-based optimisation, Recursive filtering, least squares methods (in the noiseless case), and the expectation maximization (EM) algorithm (described in greater detail in [123]).

6.2 Introducing Nonstationary concepts

In [93], an autoregressive-autoregressive (AR-AR) model is used rather than an ARMA model, that is, the observed signal is represented by an AR source passing through an AR filter (in this case the room acoustic response). Although this does not translate into a typical blur model when reformulated for two dimensional signals, it does provide a useful conceptual starting point for investigating the ideas of nonstationarity, as will be seen in Chapter 7. The key idea behind the nonstationary model used in [93] is that by allowing the source and channel models to have different rates of nonstationarity, there is no ambiguity and the parameters of each may be uniquely identified. Clearly if the source and channel parameters occupy a common space (as poles of the full transfer function in this case) then the problem comes down to assigning them correctly to the source or channel. If they do not vary with time, or vary at the same rate, clearly there is no way of telling them apart. This is demonstrated in [93] via a histogram technique: if some of the poles are time-varying then averaging them over time should leave only the stationary ones. The source model in [93] assumes that the signals are locally stationary, given a small enough time window. This results in a block-stationary AR (BSAR) model, where the signal is partitioned into regular blocks, and within each an AR model is imposed. This model will be adapted for images in Chapter 7.

Several issues have to be addressed in transferring these models to 2D signals, including causality of the model, since images do not have a natural causal ordering; representation of the PSF; inversion of the model; quantity of data (there are far fewer samples in each region than for audio), computational issues due to the "curse of dimensionality" means processing and memory requirements increase exponentially; and choice of regions in two dimensions — the shape and not just length is now an important consideration. The nonstationary models that will be considered in the remaining chapters are now developed, firstly in an abstract synthetic manner, followed by a formal specification.

6.2.1 Nonstationary Image Concepts

Assume that we have some segmentation of the image into different regions (this will be defined formally in the next section). Consider now extending the stationary ARMA model to include a different excitation variance in each region; $\mathbf{Q}_v$ may now be any diagonal matrix, parameterised by $\boldsymbol{\sigma}_v$. This gives a nonstationary variance model, as shown in fig. 6.2. This model gives a slightly more flexible image representation, and will be used as the starting point

for the experiments in Chapter 9. This model will then be extended to the one shown in fig. 6.3, where the AR parameters themselves are allowed to vary between blocks. This model can then adapt to different textures and features in the image. This model will also be used with an additional mean in each block, as in fig. 6.4, because real images are not zero mean and tend to have a local positive mean value in different parts of the image.

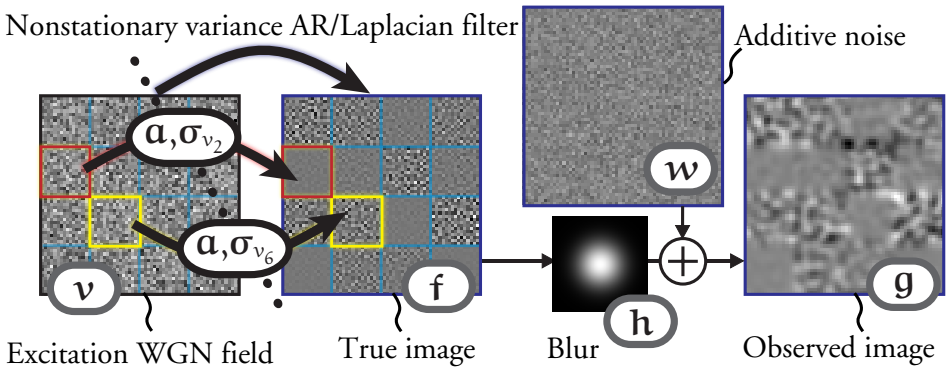


Figure 6.2: *Allowing the image model to have a space-variant variances,* $\sigma^2_{v_r}$

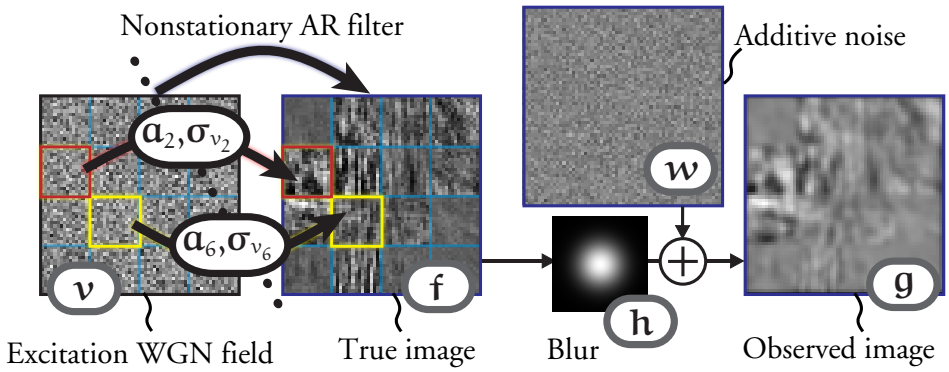


Figure 6.3: *Allowing the image model to have a space-variant variances,* $\sigma^2_{v_r}$ *and AR parameters* $\mathbf{a}_{f_r}$

6.2.2 Application to modelling of a real image

Clearly the synthetic realisations of the stochastic processes do not resemble real images; however the extra nonstationary features do allow them to better model real images and therefore act as priors. If we consider the process in reverse, we can look at the properties of the excitation signal obtained from the true image under the relation $\hat{\mathbf{v}} = (\mathbf{I} - \hat{\mathbf{A}})\mathbf{f}$. Here a ML estimate of the AR parameters, $\hat{\mathbf{a}}_f = (\mathbf{F}_a^T \mathbf{F}_a)^{-1} \mathbf{F}_a^T \mathbf{f}$ is used. Ideally this excitation signal should be white

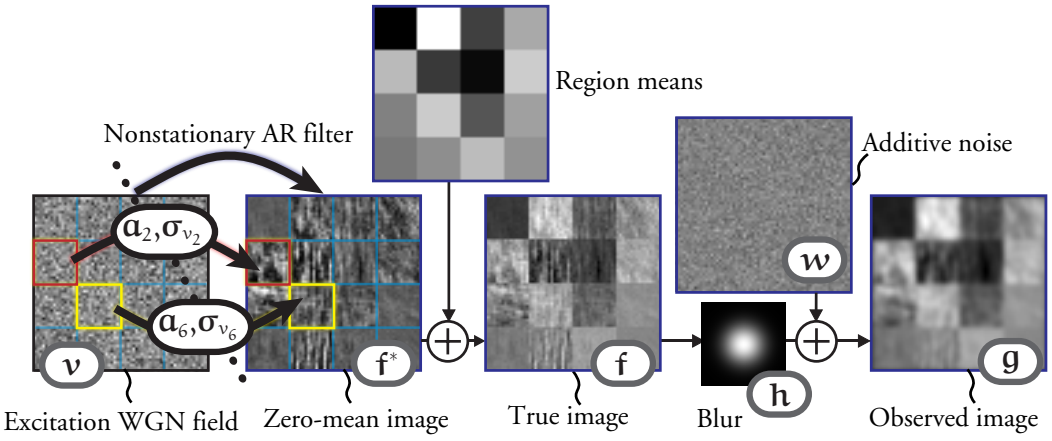


Figure 6.4: *Allowing the image model to have a space-variant variances, $\sigma^2_{v_r}$ and AR parameters $\mathbf{a}_{f_r}$, and additional local mean components*

noise if the model has done a good job in representing the image.

In fig. 6.5, beginning with the simple stationary SAR model (fixed AR parameters $\mathbf{C} = (\mathbf{I} - \mathbf{A})$ where $\mathbf{C}$ is the discrete Laplacian) in fig. 6.5(a), we see that the estimated excitation signal has large impulsive features around the edges in the original image. Allowing for nonstationary variances makes things slightly more homogeneous, as we see if we scale by the estimated variance in each block, as shown in (fig. 6.5(b)).

However the best results are obtained when we allow the AR coefficients to be estimated locally, as in fig. 6.5(c) and 6.5(d). Here we can see the advantage brought by using higher-order correlations under the AR model (5th order used), compared to more simple nonstationary processes. Ideally the image model should account for structured components in the signal, and leave behind uncorrelated noise components. Including means in the model also helps, especially for lower order models, although the main benefit comes from the AR parameters; we will see the means help bring stability to the model.

The excitation signal $\boldsymbol{v}$ should be mostly decorrelated through $\mathbf{C}$, though not necessarily independent and identically distributed (iid) Gaussian. Ideally the scaled variable $\mathbf{Q}_v^{-\frac{1}{2}}\boldsymbol{v}$ should approach a white Gaussian. We can observe the histograms of of these variables taken from the sample images. The histograms corresponding to each image in fig. 6.5 are shown in fig. 6.6, along with Gaussians of the same variance as the sample. Clearly the later figures start to better approach the Gaussian distribution, especially for lower magnitude samples, al-

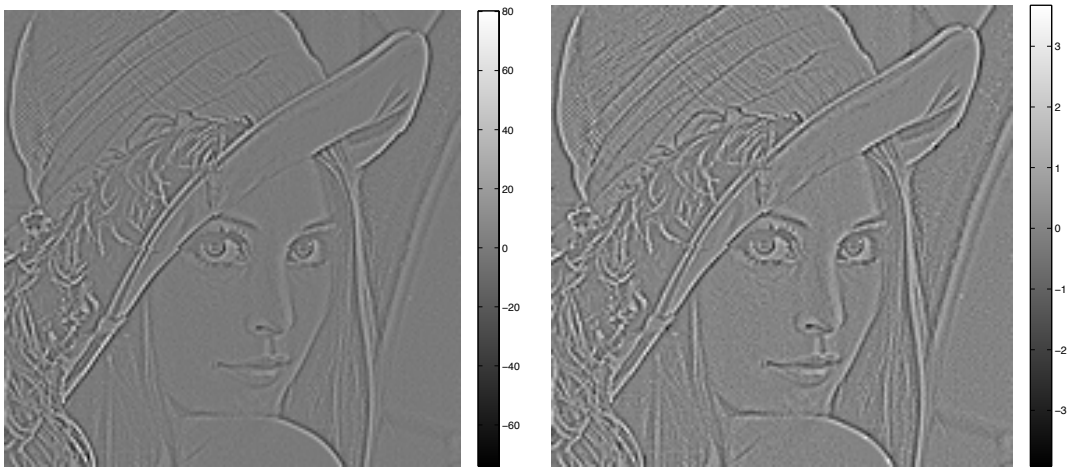
though there are obviously still slightly heavier tails in the samples where we have not entirely modelled the high frequencies in the edge transitions.

Note that whilst the locations of the edges, which contain the most energy in the estimated excitation, are spatially sparse and so could also be modelled by a heavy-tailed distribution, this will likely result in an intractable model for inference. By using a hierarchy of standard conjugate exponential distributions (Gaussian and Inverse Gamma) these problems are avoided. Indeed, an interpretation of the nonstationary variance model, which has been used by Tzikas *et al.*[217], is that the variances are modelled as a Student-t distribution. By integrating out the variances, treating them as hidden parameters in the model we obtain this result. This is effectively the same result obtained in Chapter 7, when we integrate out the variances as nuisance parameters, along with the AR model parameters.

6.2.3 Additional model extensions

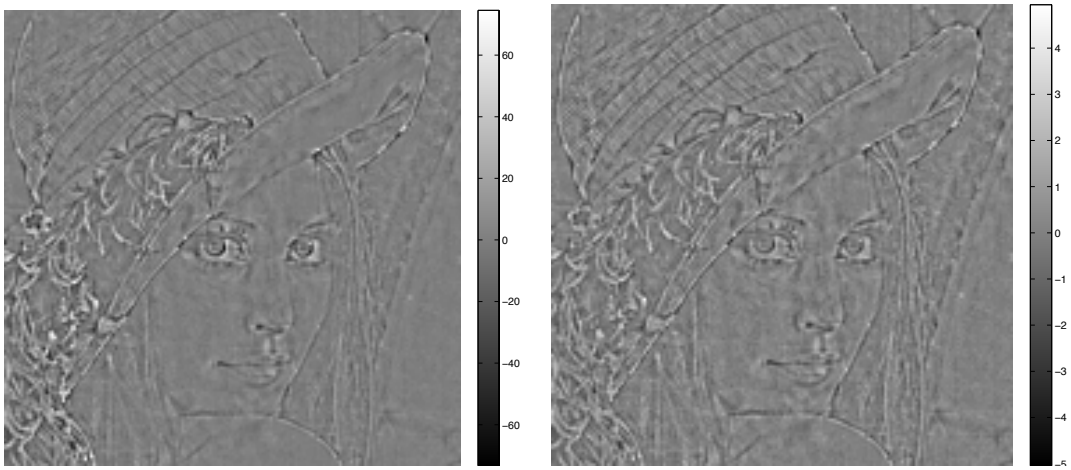
Of course we are faced with the possible problem of model over-fitting whereby too many parameters are used to represent the image; for this reason, it is likely that there will be a optimal range of AR support sizes that model the signal just well enough. However because of the hierarchical formulation and the use of Bayesian marginalisation, we will in general not simply estimate a particular set of AR parameters for an image, but attempt to consider a range of their likely values. The AR model tends to introduce parameters that model low to mid frequencies which tends to make it suitable as a prior; it is an approximation of the image based on the observations. Moreover in the methods in Chapter 9 and 10, the AR parameters are effectively only used for regularising the image restorations. Therefore the issue of over-fitting should not bias the results for reasonable support sizes.

Furthermore, the assumption of regular rectangular blocks in the image is clearly not a particularly realistic one. In terms of fitting a model for locally approximating the image (as in Chapter 7) this is not a huge problem; though if the model is used as a prior it may introduce artefacts at the boundaries. Clearly one solution to this is to also perform a segmentation of the image simultaneously as the model parameters are estimated, such that region boundaries are made to coincide with the edges in the image. Alternatively, small or single pixel blocks may be used. However, with single pixel regions, there is an issue with correlation between neighbouring blocks — one possible solution to this problem would be to use a prior that correlates the variances, however we leave this to future work. The method that is used in



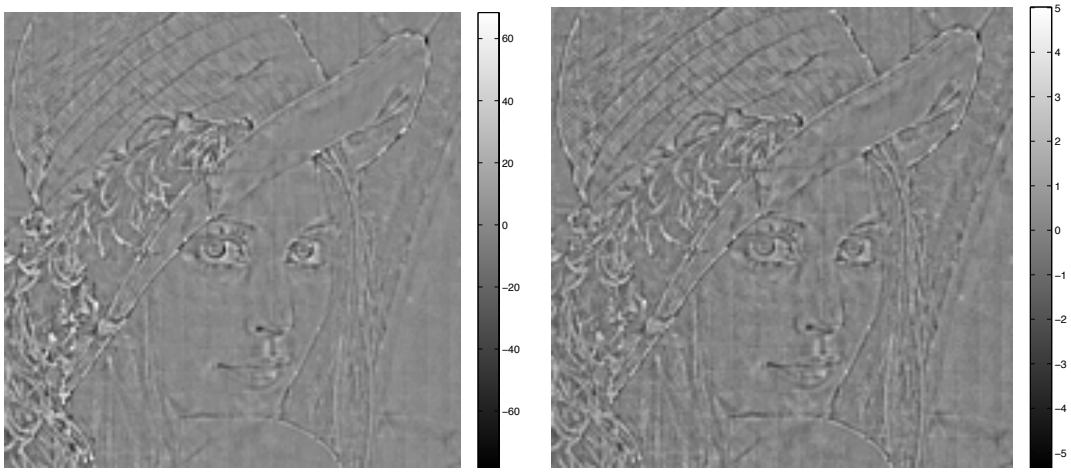
(a) $\boldsymbol{\nu}$ with $\mathbf{C}$ =Laplacian

(b) $\mathbf{Q}_{\nu}^{-\frac{1}{2}}\boldsymbol{\nu}$ with $\mathbf{C}$ =Laplacian



(c) $\boldsymbol{\nu}$ with AR estimation

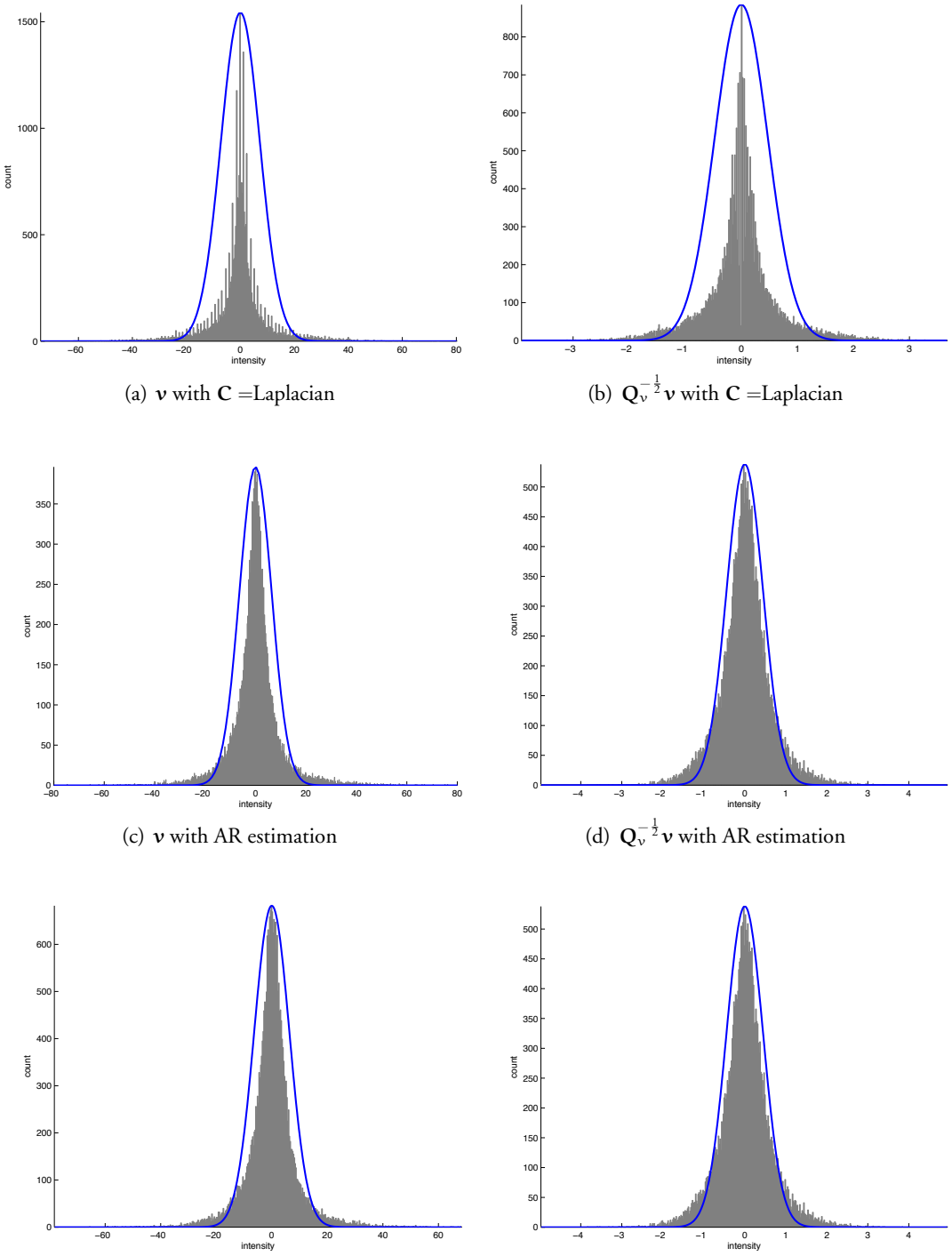
(d) $\mathbf{Q}_{\nu}^{-\frac{1}{2}}\boldsymbol{\nu}$ with AR estimation



(e) $\boldsymbol{\nu}$ with AR & means estimation

(f) $\mathbf{Q}_{\nu}^{-\frac{1}{2}}\boldsymbol{\nu}$ with AR & means estimation

Figure 6.5: *Excitation signal estimated from a real image under different models. Left column shows the excitation signal* $\boldsymbol{\nu}$*, right column shows after scaling by the standard deviation of each* 8×8 *block.*



(a) $\boldsymbol{\nu}$ with $\mathbf{C}$ =Laplacian

(b) $\mathbf{Q}_{\nu}^{-\frac{1}{2}}\boldsymbol{\nu}$ with $\mathbf{C}$ =Laplacian

(c) $\boldsymbol{\nu}$ with AR estimation

(d) $\mathbf{Q}_{\nu}^{-\frac{1}{2}}\boldsymbol{\nu}$ with AR estimation

(e) $\boldsymbol{\nu}$ with AR & means estimation

(f) $\mathbf{Q}_{\nu}^{-\frac{1}{2}}\boldsymbol{\nu}$ with AR & means estimation

Figure 6.6: *Histograms of excitation signal estimated from a real image under different models. Left column shows the excitation signal* $\boldsymbol{\nu}$*, right column shows after scaling by the standard deviation of each* 8×8 *block. Note the better Gaussianity under the AR models.*

Chapter 9 and 10 for mitigating block boundary artefacts is in fact much simpler, and relies on estimating the parameters for different random block grids at each iteration. This will be discussed further in §9.5.1.

6.3 Nonstationary Image Model

We will now present the mathematical formulation for the basic generative image model that forms the main component of the hierarchical model. This is used as the prior for $\mathbf{f}$, and also represents a general and flexible 2D signal model that may be adapted to represent the intensity values of the PSF .

6.3.1 Segmentation of the canvas

We begin by considering an image $\mathbf{f}$, defined on the canvas $\mathcal{S}_f$, of size L_+ pixels. We segment this canvas into R blocks or regions, $\mathcal{R}_r \subset \mathcal{S}_f$ where $r \in \{1 \cdots R\}$, each of size L_r pixels (the size of each block may vary in general). We denote by $\mathcal{R}_{r_+}$ the region which is a dilation of $\mathcal{R}_r$ by a support region of shape $\mathcal{S}_{a_f}$. The difference between these two regions, which contains pixels bordering $\mathcal{R}_r$ that are in $\mathcal{R}_{r_+}$ is denoted as $\mathcal{R}_{r_b}$, such that $\mathcal{R}_{r_+} = \mathcal{R}_r \cup \mathcal{R}_{r_b}$ (see Fig. 6.7). Within each region, we assume a common homogeneous image formation model. A fixed segmentation of the image is used, but in principle the blind image restoration (BIR) and segmentation problems may be tackled jointly with the same model using the Bayesian framework.

Under the fixed scheme we simply use regular square blocks of size $m_r \times n_r$ pixels, such that L_r is the same for all regions.[1] The model described however applies equally to any other given segmentation, and may even be simply extended to estimate the segmentation as part of the model. Essentially, we may think of the segmentation as a hidden labelling variable Ψ, which the model is conditional on. For now, we simply hold Ψ fixed.

The principle of the BSAR model is to find a set of coefficients $\boldsymbol{a}_{f_r}$ that model the image in each local region $\mathcal{R}_r$, such that each pixel is the sum of an excitation signal $v_r(i,j)$ and the weighted combination of the pixel's neighbours within a support region of shape $\mathcal{S}_{a_f}$, assumed to be square, whose size is $L_{a_f} = (2P+1)^2 - 1$ for the non-causal model (or $(P+1)^2 - 1$ in the

[1] Apart from at the image boundaries, where the grid may not intersect exactly with the image and smaller rectangular blocks will then be used.

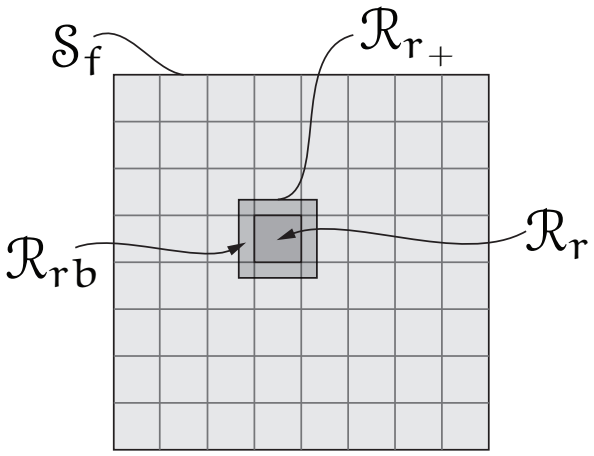


Figure 6.7: *A regular segmentation of the image support $\mathcal{S}_f$, with the r^{th} region $\mathcal{R}_r$, its border pixels $\mathcal{R}_{r_b}$, and their union $\mathcal{R}_{r_+}$ shown.*

causal case), where P is the AR model order. If the model is a good fit, the excitation signal, or modelling error, will be small and uncorrelated with the image, and resemble WGN, as has been described in the previous section.

6.3.2 Image vectors in each region

For a given region $\mathcal{R}_r$, the image values inside it are denoted $f_r(i,j)$, and these are raster scanned to give the vector $\mathbf{f}_r$ (note that this is possible even for irregularly shaped regions).

The signal at the interior boundary of $\mathcal{R}_r$ depends on its neighbours outside of the region, according to the extent of $\mathcal{S}_{a_f}$. These exterior pixels, which are in $\mathcal{R}_{r_b}$, come from the neighbouring regions of $\mathcal{R}_r$, and are vectorised as $\mathbf{f}_{rb}$.

The vector containing all pixels inside $\mathcal{R}_{r_+}$, is $\mathbf{f}_{r_+} = [\mathbf{f}_r^T, \mathbf{f}_{rb}^T]^T$. We will write the length of $\mathbf{f}_{rb}$ as L_{rb} and of $\mathbf{f}_{r_+}$ as L_{r_+}.

6.3.3 2D AR model in each region

We now proceed to model the signal in each block. Let us assume for the moment that the image in $\mathcal{R}_r$ is a zero mean signal, which we will denote $f_r^*(i,j)$.

Within each region, we define a P^{th} order AR model according to:

$$f_r^*(i,j) = \sum_{\forall(k,l \in \mathcal{S}_{a_f})} f_r^*(i-k, j-l) a_r(k,l) + \nu_r(i,j) \tag{6.7}$$

where and $\nu_r(i,j) \sim \mathcal{N}(0, \sigma_{\nu_r}^2)$ is a zero-mean WGN excitation signal representing the local image activity. The AR coefficients a_r are non-zero within a $(2P+1) \times (2P+1)$ non-causal support region $\mathcal{S}_{a_f}$, excluding the pixel $k, l = 0$.

In vector form, after lexicographic ordering, (6.7) is written as

$$\mathbf{f}_r^* = \mathbf{A}_{f_{r_+}} \mathbf{f}_{r_+}^* + \boldsymbol{\nu}_r \tag{6.8}$$

where $\boldsymbol{\nu}_r \sim \mathcal{N}(\mathbf{0}, \sigma_{\nu_r}^2 \mathbf{I}_{L_r})$ is the excitation in $\mathcal{R}_r$.[2] $\mathbf{A}_{f_{r_+}}$ is a $L_r \times L_{r_+}$ matrix that performs convolution between $a_r(k,l)$ and an image defined on $\mathcal{R}_{r_+}$, giving an output on $\mathcal{R}_r$. We can partition this as $\mathbf{A}_{f_{r_+}} = [\mathbf{A}_{f_r} | \mathbf{A}_{f_{rb}}]$, where $\mathbf{A}_{f_r}$ takes input only from $\mathcal{R}_r$ and $\mathbf{A}_{f_{rb}}$ from $\mathcal{R}_{r_b}$. Note that while $\mathbf{A}_{f_r}$ is BTTB, $\mathbf{A}_{f_{r_+}}$ is only BTTB under suitable re-ordering. Now defining $\mathbf{I}_{r_+} = [\mathbf{I}_{L_r} | \mathbf{0}_{L_r \times L_{rb}}]$, the excitation signal in $\mathcal{R}_r$ may be written

$$\boldsymbol{\nu}_r = (\mathbf{I}_{r_+} - \mathbf{A}_{f_{r_+}}) \mathbf{f}_{r_+}^* = \mathbf{C}_{f_{r_+}} \mathbf{f}_{r_+}^* \tag{6.9}$$

or equivalently

$$\boldsymbol{\nu}_r = \mathbf{f}_{r_+}^* - \mathbf{F}_{a_r}^* \mathbf{a}_{f_r}. \tag{6.10}$$

The vector $\mathbf{a}_{f_r}$ is of length L_{a_f}. The $L_r \times L_{a_f}$ matrix $\mathbf{F}_{a_r}^*$ is also BTTB, and performs convolution between $\mathbf{f}_{r_+}^*$ and AR coefficients defined on $\mathcal{S}_{a_f}$. We can also expand $\mathbf{C}_{f_{r_+}} \mathbf{f}_{r_+}^*$ as $\mathbf{C}_{f_r} \mathbf{f}_r^* - \mathbf{A}_{f_{rb}} \mathbf{f}_{rb}^*$, where $\mathbf{C}_{f_r} \triangleq (\mathbf{I}_{L_r} - \mathbf{A}_{f_r})$. Observe that the vectors $\mathbf{F}_{a_r}^*$ and $\mathbf{f}_{r_+}^*$ depend on pixels from the neighbouring regions to $\mathcal{R}_r$.

6.3.4 Whole image model

Because the excitation is independent across blocks, we can use it to form a common model for the whole image. Define $\boldsymbol{\nu}' = [\boldsymbol{\nu}_1^T, \cdots, \boldsymbol{\nu}_R^T]^T$, the concatenation of the excitation across all the regions in the canvas $\mathcal{S}_f$, and a similar block-wise lexicographic scanning of $f(i,j)$ as

[2] We use the notation $\mathbf{1}_M$ or $\mathbf{0}_M$ for a vector of M ones or zeros, and $\mathbf{1}_{M \times N}$ or $\mathbf{0}_{M \times N}$ for an $M \times N$ matrix of ones or zeros. $\mathbf{I}_M$ is the $M \times M$ identity matrix.

$\mathbf{f}'^{*} = [\mathbf{f}_1^{*\mathsf{T}}, \cdots, \mathbf{f}_R^{*\mathsf{T}}]^{\mathsf{T}}$. Then by stacking (6.9) and (6.10) for all $r = 1 \cdots R$ we may write

$$\boldsymbol{\nu}' = (\mathbf{I}_{L_+} - \mathbf{A}_f')\mathbf{f}'^{*} = \mathbf{C}_f'\mathbf{f}'^{*} \tag{6.11}$$

$$= \mathbf{f}'^{*} - \mathbf{F}_a'^{*}\boldsymbol{\alpha}_f, \tag{6.12}$$

where we have also defined $\boldsymbol{\alpha}_f = [\boldsymbol{\alpha}_{\mathbf{f}_1}{}^{\mathsf{T}}, \cdots, \boldsymbol{\alpha}_{\mathbf{f}_R}{}^{\mathsf{T}}]^{\mathsf{T}}$. Here $\mathbf{F}_a'^{*}$ is a block-diagonal matrix with blocks $\mathbf{F}_{a_r}^{*}$ for $r = 1 \cdots R$. The matrices $\mathbf{A}_f'$ and $\mathbf{C}_f'$ are similarly constructed with their diagonal blocks as $\mathbf{A}_{f_r}$ and $\mathbf{C}_{f_r}$, but off-diagonal blocks contain appropriate parts from $\mathbf{A}_{f_{rb}}$ according to the neighbouring regions, for $r = 1 \cdots R$.

The vector $\boldsymbol{\nu}'$ is not in standard lexicographic ordering. We can re-order it as $\boldsymbol{\nu} = \mathbf{P}\boldsymbol{\nu}'$, where $\mathbf{P}$ is a permutation matrix that converts the block-wise scanning to lexicographic ordering across $\mathcal{S}_f$. Then we have

$$\boldsymbol{\nu} = (\mathbf{I}_{L_+} - \mathbf{A}_f)\mathbf{f}^{*} = \mathbf{C}_f\mathbf{f}^{*} \tag{6.13}$$

$$= \mathbf{f}^{*} - \mathbf{F}_a^{*}\boldsymbol{\alpha}_f \tag{6.14}$$

where $\mathbf{C}_f$ and $\mathbf{A}_f$ are nonstationary BTTB matrices with entries taken from the matrices $\mathbf{C}_{f_{r_+}}$ and $\mathbf{A}_{f_{r_+}}$ for $r = 1 \cdots R$, relating pixels in $\boldsymbol{\nu}$ and $\mathbf{f}$ according to (6.9). $\mathbf{F}_a^{*}$ contains entries from $\mathbf{F}_{a_r}^{*}$ according to (6.10).

Now, using the PDF of the excitation signal in each block $\boldsymbol{\nu}_r \sim \mathcal{N}(\mathbf{0}, \sigma_{\nu_r}^2 \mathbf{I}_{L_r})$, and the fact that these are independent, we may write

$$p\left(\boldsymbol{\nu}' \,|\, \boldsymbol{\sigma}_\nu\right) = \prod_{r \in \{1 \cdots R\}} p\left(\boldsymbol{\nu}_r \,\middle|\, \sigma_{\nu_r}^2\right) \tag{6.15}$$

$$= \prod_{r \in \{1 \cdots R\}} (2\pi)^{-\frac{L_r}{2}} \sigma_{\nu_r}^{-L_r} \exp\left[-\frac{1}{2}\sigma_{\nu_r}^{-2}\boldsymbol{\nu}_r^{\mathsf{T}}\boldsymbol{\nu}_r\right] \tag{6.16}$$

$$= (2\pi)^{-L_+} \det|\mathbf{Q}_{\nu'}|^{-1} \exp\left[-\frac{1}{2}\boldsymbol{\nu}'^{\mathsf{T}}\mathbf{Q}_{\nu'}^{-1}\boldsymbol{\nu}'\right] \tag{6.17}$$

where $\boldsymbol{\sigma}_\nu = [\sigma_{\nu_1}^2, \cdots, \sigma_{\nu_R}^2]^{\mathsf{T}}$, and the diagonal matrix $\mathbf{Q}_{\nu'}$ is formed from the excitation variances in each block, $\sigma_{\nu_r}^2$; for example as $\mathbf{Q}_{\nu'} = \mathrm{diag}(\boldsymbol{\sigma}_\nu) \otimes \mathbf{I}_{L_r}$, in the case when the blocks are uniform size L_r pixels, or in general $\mathbf{Q}_{\nu'} = \mathrm{diag}([\sigma_{\nu_1}^2 \mathbf{1}_{L_1}^{\mathsf{T}} \cdots \sigma_{\nu_R}^2 \mathbf{1}_{L_R}^{\mathsf{T}}]^{\mathsf{T}})$. Now the PDF of $\boldsymbol{\nu} = \mathbf{P}\boldsymbol{\nu}'$ is given by

$$p_{\boldsymbol{\nu}}\left(\boldsymbol{\nu} \,|\, \boldsymbol{\sigma}_\nu\right) = p_{\boldsymbol{\nu}'}\left(\mathbf{P}\boldsymbol{\nu}' \,|\, \boldsymbol{\sigma}_\nu\right) \tag{6.18}$$

$$= (2\pi)^{-L_+} \det|\mathbf{P}\mathbf{Q}_{\nu'}\mathbf{P}^T|^{-1} \exp\left[-\frac{1}{2}\boldsymbol{\nu}^T\mathbf{P}^T\mathbf{Q}_{\nu'}^{-1}\mathbf{P}\boldsymbol{\nu}'\right] \tag{6.19}$$

$$= (2\pi)^{-L_+} \det|\mathbf{Q}_\nu|^{-1} \exp\left[-\frac{1}{2}\boldsymbol{\nu}^T\mathbf{Q}_\nu^{-1}\boldsymbol{\nu}\right] \tag{6.20}$$

where the covariance of $\boldsymbol{\nu}$ is defined as the diagonal matrix $\mathbf{Q}_\nu = \mathbf{P}\mathbf{Q}_{\nu'}\mathbf{P}^T$.

Finally, by applying a probability transformation to (6.20), using (6.13), the extended zero-mean image prior PDF may be written as

$$\begin{aligned} p(\mathbf{f}^* \mid \boldsymbol{\alpha}_f, \sigma_\nu) &= (2\pi)^{-L_+} \det|\boldsymbol{\Sigma}_{f_+}^{-1}| \cdot \\ &\exp\left[-\frac{1}{2}(\mathbf{C}_f\mathbf{f}^*)^T\mathbf{Q}_\nu^{-1}(\mathbf{C}_f\mathbf{f}^*)\right] \end{aligned} \tag{6.21}$$

$$= \mathcal{N}\left(\mathbf{f}^* \,\middle|\, \mathbf{0}_{L_+}, \boldsymbol{\Sigma}_{f_+}\right), \tag{6.22}$$

where
$$\boldsymbol{\Sigma}_{f_+} = \mathbb{E}\left[\mathbf{f}^*\mathbf{f}^{*T}\right] = \mathbf{C}_f^{-1}\mathbf{Q}_\nu\mathbf{C}_f^{-T} \tag{6.23}$$

Note that the while the excitation signal $\boldsymbol{\nu}$ is independent in each region, the image in each region depends on its neighbours.

In summary, the block stationary model splits the image into blocks within which the image is assumed a stationary process, allowing different AR parameters in each block. With all the AR parameters organised into a vector $\boldsymbol{\alpha}_f$, each row of the $\mathbf{F}_a^*$ matrix must contain the contributions from other pixels in the image in the column positions corresponding to the position of the particular AR parameters from the block to which this pixel in the same row of $\mathbf{f}^*$ belongs. The width of the $\mathbf{F}_a^*$ matrix depends on the number of AR parameters used in the model, which again depends on the model order and whether the source is assumed causal or non-causal (i.e., by the size of the support region, S_a).

6.4 Extension to include means

So far, we assumed that the image was a zero-mean signal. In practice, unlike e.g. audio signals, it will be positive valued, with a spatially varying local mean. Therefore we can add a mean value to each region to improve the model. We let the zero mean signal $f_r^*(i,j)$ be equal to $f_r(i,j) - \mu_{f_r}$, where μ_{f_r} is the mean of region r. The results from before hold, but we should be more careful in defining the excitation signals to find the PDF of $\mathbf{f}$. The details are presented in this section.

6.4.1 2D AR model with means in each region

We define the mean vector for the regions as $\mathbf{m}_{f_r} = \mu_{f_r}\mathbf{1}_{L_r}$, $\mathbf{m}_{f_{r_+}} = \mu_{f_r}\mathbf{1}_{L_{r_+}}$, $\mathbf{m}_{f_{rb}} = \mu_{f_r}\mathbf{1}_{L_{rb}}$. Note the complication here compared to the case without the mean, that in considering each region $\mathcal{R}_{r_+}$ we consider the signal to have the same mean as inside $\mathcal{R}_r$. Then, letting $\mathbf{f}_r^* = \mathbf{f}_r - \mathbf{m}_{f_r}$, the vector form of (6.7) becomes

$$\mathbf{f}_r - \mathbf{m}_{f_r} = \mathbf{A}_{f_{r_+}}(\mathbf{f}_{r_+} - \mathbf{m}_{f_{r_+}}) + \boldsymbol{\nu}_r \tag{6.24}$$

$$= \mathbf{A}_{f_r}(\mathbf{f}_r - \mathbf{m}_{f_r}) + \mathbf{A}_{f_{rb}}(\mathbf{f}_{rb} - \mathbf{m}_{f_{rb}}) + \boldsymbol{\nu}_r \tag{6.25}$$

$$= \mathbf{A}_{f_{r_+}}\mathbf{f}_{r_+} - \mathbf{A}_{m_r}\mathbf{m}_{f_r} + \boldsymbol{\nu}_r \tag{6.26}$$

We have used a trick in (6.26) to rewrite the term $\mathbf{A}_{f_{r_+}}\mathbf{m}_{f_{r_+}} = \mathbf{A}_{f_r}\mathbf{m}_{f_r} + \mathbf{A}_{f_{rb}}\mathbf{m}_{f_{rb}}$ as $\mathbf{A}_{m_r}\mathbf{m}_{f_r}$. Recall that $\mathbf{A}_{f_r}$ is an $L_r \times L_r$ BTTB matrix, and $\mathbf{A}_{f_{r_+}}$ extends this convolution to include input from $\mathcal{R}_{r_b}$. We then define $\mathbf{A}_{m_r}$ as an $L_r \times L_r$ BCCB convolution matrix; it performs convolution between $a_r(k,l)$ and an image defined on $\mathcal{R}_r$ under toroidal boundary conditions. Then the above equivalence holds, due to the fact that the mean vectors have the same value across $\mathcal{R}_{r_+}$.

An example will clarify this; we show a 1D case for simplicity but the same result holds for 2D images under lexicographic ordering. The equivalence is expanded as:

$$
\overbrace{\left[\begin{array}{cccccccccc|ccccccc}
 & & & & & & & & & & & & & & & & \\
0 & a_{-1} & \dots & a_{-P} & 0 & \dots & 0 & 0 & \dots & 0 & a_P & \dots & a_1 & 0 & \dots & 0 \\
a_1 & 0 & a_{-1} & \dots & a_{-P} & \ddots & & \ddots & \ddots & \vdots & 0 & \ddots & \vdots & \vdots & & \vdots \\
\vdots & a_1 & \ddots & \ddots & & \ddots & \ddots & & \ddots & 0 & \vdots & \ddots & a_P & \vdots & & \vdots \\
a_P & \vdots & \ddots & \ddots & \ddots & & \ddots & \ddots & & 0 & \vdots & & 0 & \vdots & & \vdots \\
0 & a_P & & \ddots & \ddots & \ddots & & \ddots & \ddots & \vdots & \vdots & & \vdots & \vdots & & \vdots \\
\vdots & \ddots & \ddots & & \ddots & \ddots & \ddots & & \ddots & 0 & \vdots & & \vdots & \vdots & & \vdots \\
0 & & \ddots & \ddots & & \ddots & \ddots & \ddots & & a_{-P} & \vdots & & \vdots & 0 & & \vdots \\
0 & \ddots & & \ddots & \ddots & & \ddots & \ddots & \ddots & \vdots & \vdots & & \vdots & a_{-P} & \ddots & \vdots \\
\vdots & \ddots & \ddots & & \ddots & \ddots & & \ddots & \ddots & a_{-1} & \vdots & & \vdots & \vdots & \ddots & 0 \\
\underbrace{0 \quad \dots \quad 0 \quad 0 \quad \dots \quad 0 \quad a_P \quad \dots \quad a_1 \quad 0}_{\mathbf{A}_{f_r}} & & & & & & & & & & \underbrace{0 \quad \dots \quad 0 \quad a_1 \quad \dots \quad a_{-P}}_{\mathbf{A}_{f_{rb}}} & & & & & &
\end{array}\right]}^{\mathbf{A}_{f_{r_+}}}
\overbrace{\left[\begin{array}{c}
\mu_{f_r} \\
\vdots \\
\vdots \\
\vdots \\
\vdots \\
\vdots \\
\vdots \\
\vdots \\
\vdots \\
\vdots \\
\mu_{f_r}
\end{array}\right]}^{\mathbf{m}_{f_{r_+}}}
\tag{6.27}
$$

$$
= \overbrace{\left[\begin{array}{cccccccccc}
0 & a_{-1} & \dots & a_{-P} & 0 & \dots & 0 & a_P & \dots & a_1 \\
a_1 & 0 & a_{-1} & \dots & a_{-P} & \ddots & & \ddots & \ddots & \vdots \\
\vdots & a_1 & \ddots & \ddots & & \ddots & \ddots & & \ddots & a_P \\
a_P & \vdots & \ddots & \ddots & \ddots & & \ddots & \ddots & & 0 \\
0 & a_P & & \ddots & \ddots & \ddots & & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & & \ddots & \ddots & \ddots & & \ddots & 0 \\
0 & & \ddots & \ddots & & \ddots & \ddots & \ddots & & a_{-P} \\
a_{-P} & \ddots & & \ddots & \ddots & & \ddots & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & & \ddots & \ddots & & \ddots & \ddots & a_{-1} \\
a_1 & \dots & a_{-P} & 0 & \dots & 0 & a_P & \dots & a_1 & 0
\end{array}\right]}^{\mathbf{A}_{m_r}}
\overbrace{\left[\begin{array}{c}
\mu_{f_r} \\
\vdots \\
\vdots \\
\vdots \\
\vdots \\
\vdots \\
\vdots \\
\vdots \\
\vdots \\
\vdots \\
\mu_{f_r}
\end{array}\right]}^{\mathbf{m}_{f_r}}
\tag{6.28}
$$

Then the excitation signal in $\mathcal{R}_r$ may be written

$$
\boldsymbol{\nu}_r = (\mathbf{I}_{r_+} - \mathbf{A}_{f_{r_+}})(\mathbf{f}_{r_+} - \mathbf{m}_{f_{r_+}}) \tag{6.29}
$$

$$= \mathbf{C}_{\mathfrak{f}_{r_+}} \mathfrak{f}_{r_+} - \boldsymbol{\eta}_{\mathfrak{f}_r} \tag{6.30}$$

or equivalently $$\boldsymbol{\nu}_r = (\mathfrak{f}_r - \mathfrak{m}_{\mathfrak{f}_r}) - (\mathbf{F}_{\mathfrak{a}_r} - \mathbf{M}_{\mathfrak{a}_r})\mathfrak{a}_{\mathfrak{f}_r}. \tag{6.31}$$

where $\mathbf{M}_{\mathfrak{a}_r} = \mu_{\mathfrak{f}_r} \mathbf{1}_{L_r \times L_a}$, and we introduce the vector $\boldsymbol{\eta}_{\mathfrak{f}_r}$, which may be expanded as either $\mathbf{C}_{\mathfrak{f}_{r_+}} \mathfrak{m}_{\mathfrak{f}_{r_+}}$ or equivalently $\mathbf{C}_{\mu_r} \mathfrak{m}_{\mathfrak{f}_r}$, where $\mathbf{C}_{\mu_r} \triangleq (\mathbf{I}_{L_r} - \mathbf{A}_{\mathfrak{m}_r})$. The $L_r \times L_{\mathfrak{a}_f}$ matrix $\mathbf{F}_{\mathfrak{a}_r} = \mathbf{F}^*_{\mathfrak{a}_r} + \mathbf{M}_{\mathfrak{a}_r}$ performs convolution between $\mathfrak{f}_{r_+}$ and AR coefficients defined on $\mathcal{S}_{\mathfrak{a}_f}$. Finally note again that we can also expand $\mathbf{C}_{\mathfrak{f}_{r_+}} \mathfrak{f}_{r_+}$ as $\mathbf{C}_{\mathfrak{f}_r} \mathfrak{f}_r - \mathbf{A}_{\mathfrak{f}_{rb}} \mathfrak{f}_{rb}$. This is useful to help to define the complete image model.

6.4.2 Whole image model with means

If we now define the stacked vectors $\mathfrak{f}' = [\mathfrak{f}_1^T, \cdots, \mathfrak{f}_R^T]^T$, $\mathfrak{m}_\mathfrak{f}{}' = [\mathfrak{m}_{\mathfrak{f}_1}^T, \cdots, \mathfrak{m}_{\mathfrak{f}_R}^T]^T$ and $\boldsymbol{\eta}'_\mathfrak{f} = \boldsymbol{\eta}_{\mathfrak{f}_1}^T, \cdots, \boldsymbol{\eta}_{\mathfrak{f}_R}^T]^T$, and concatenate $\boldsymbol{\nu}' = [\boldsymbol{\nu}_1^T, \cdots, \boldsymbol{\nu}_R^T]^T$ as before, we have

$$\boldsymbol{\nu}' = (\mathbf{I}_{L_+} - \mathbf{A}'_\mathfrak{f})\mathfrak{f}' - (\mathbf{I}_{L_+} - \mathbf{A}'_\mathfrak{m})\mathfrak{m}_\mathfrak{f}{}' \tag{6.32}$$

$$= \mathbf{C}'_\mathfrak{f}\mathfrak{f}' - \boldsymbol{\eta}'_\mathfrak{f} \tag{6.33}$$

or equivalently $$\boldsymbol{\nu} = (\mathfrak{f}' - \mathfrak{m}_\mathfrak{f}{}') - (\mathbf{F}'_\mathfrak{a} - \mathbf{M}'_\mathfrak{a})\mathfrak{a}_\mathfrak{f}, \tag{6.34}$$

where we can also write $\boldsymbol{\eta}'_\mathfrak{f} = \mathbf{C}'_{\mu_+}\mathfrak{m}_\mathfrak{f}{}' = (\mathbf{I}_{L_+} - \mathbf{A}'_\mathfrak{m})\mathfrak{m}_\mathfrak{f}{}'$. Here the matrices $\mathbf{F}'_\mathfrak{a}$, $\mathbf{M}'_\mathfrak{a}$, $\mathbf{C}'_{\mu_+}$, and $\mathbf{A}'_\mathfrak{m}$ are block-diagonal matrices whose blocks are the respective terms in (6.29), (6.30), and (6.31) for $r = 1 \cdots R$. The matrices $\mathbf{A}'_\mathfrak{f}$ and $\mathbf{C}'_\mathfrak{f}$ are similarly constructed with their diagonal blocks as $\mathbf{A}_{\mathfrak{f}_r}$ and $\mathbf{C}_{\mathfrak{f}_r}$, but off-diagonal blocks contain appropriate parts from $\mathbf{A}_{\mathfrak{f}_{rb}}$ according to the neighbouring regions, for $r = 1 \cdots R$.

By reordering the vectors $\mathfrak{f}'$, $\mathfrak{m}_\mathfrak{f}{}'$, $\boldsymbol{\eta}'_\mathfrak{f}$ and $\boldsymbol{\nu}'$ into the standard lexicographic scanning as before using the permutation matrix $\mathbf{P}$, the complete BSAR model with means may be written as

$$\mathfrak{f} - \mathfrak{m}_\mathfrak{f} = \mathbf{A}_\mathfrak{f}\mathfrak{f} - \mathbf{A}_\mathfrak{m}\mathfrak{m}_\mathfrak{f} + \boldsymbol{\nu}, \tag{6.35}$$

and we may rewrite the definition of $\boldsymbol{\nu}$ using the exact same terms as in (6.32), (6.33) and (6.34), without the dashes:

$$\boldsymbol{\nu} = (\mathbf{I}_{L_+} - \mathbf{A}_\mathfrak{f})\mathfrak{f} - (\mathbf{I}_{L_+} - \mathbf{A}_\mathfrak{m})\mathfrak{m}_\mathfrak{f} \tag{6.36}$$

$$= \mathbf{C}_\mathfrak{f}\mathfrak{f} - \boldsymbol{\eta}_\mathfrak{f} \tag{6.37}$$

or equivalently $$\boldsymbol{\nu} = (\mathbf{f} - \mathbf{m}_\mathbf{f}) - (\mathbf{F}_\alpha - \mathbf{M}_\alpha)\boldsymbol{\alpha}_\mathbf{f}, \tag{6.38}$$

where $\boldsymbol{\eta}_\mathbf{f} = \mathbf{C}_{\mu_+}\mathbf{m}_\mathbf{f} = (\mathbf{I}_{L_+} - \mathbf{A}_m)\mathbf{m}_\mathbf{f}$. The matrices without dashes are permuted from their dashed versions according to the new ordering. Observe that if the vectors $\boldsymbol{\alpha}_{\mathbf{f}_r}$ were all the same, then $\mathbf{C}_\mathbf{f}$ and $\mathbf{A}_\mathbf{f}$ would now have BTTB forms.

The PDF of $\boldsymbol{\nu} = \mathbf{P}\boldsymbol{\nu}'$ is as before in (6.20). Thus by applying a probability transformation to (6.20), using (6.37), the non-zero-mean extended image prior PDF may be written as

$$p(\mathbf{f} \mid \boldsymbol{\alpha}, \sigma_\nu, \mathbf{m}_\mathbf{f}) = (2\pi)^{-L_+} \det|\boldsymbol{\Sigma}_{\mathbf{f}_+}^{-1}| \exp\left[-\frac{1}{2}(\mathbf{C}_\mathbf{f}\mathbf{f} - \mathbf{C}_{\mu_+}\mathbf{m}_\mathbf{f})^T \mathbf{Q}_\nu^{-1}(\mathbf{C}_\mathbf{f}\mathbf{f} - \mathbf{C}_{\mu_+}\mathbf{m}_\mathbf{f})\right] \tag{6.39}$$

$$= \mathcal{N}\left(\mathbf{f} \,\middle|\, \boldsymbol{\mu}_{\mathbf{f}_+}, \boldsymbol{\Sigma}_{\mathbf{f}_+}\right), \tag{6.40}$$

where $$\boldsymbol{\Sigma}_{\mathbf{f}_+} = \mathbb{E}\left[(\mathbf{f} - \boldsymbol{\mu}_{\mathbf{f}_+})(\mathbf{f} - \boldsymbol{\mu}_{\mathbf{f}_+})^T\right] = (\mathbf{I}_{L_+} - \mathbf{A}_\mathbf{f})^{-1}\mathbf{Q}_\nu(\mathbf{I}_{L_+} - \mathbf{A}_\mathbf{f})^{-T}. \tag{6.41}$$

and $$\boldsymbol{\mu}_{\mathbf{f}_+} = (\mathbf{I}_{L_+} - \mathbf{A}_\mathbf{f})^{-1}(\mathbf{I}_{L_+} - \mathbf{A}_m)\mathbf{m}_\mathbf{f}. \tag{6.42}$$

Here we denote $\boldsymbol{\mu} = [\mu_{\mathbf{f}_1}, \cdots, \mu_{\mathbf{f}_R}]^T$, i.e. the vector of coefficients parameterising the mean image — rather than the mean image within $\mathcal{S}_\mathbf{f}$, which is $\boldsymbol{\mu}$. This completes the full image model specification, with one small caveat. To get rid of the dependance on $\boldsymbol{\alpha}_\mathbf{f}$ in the determinant in (6.39), we will may in practice assume a causal AR model to make some later calculations tractable; however if we assume that $\boldsymbol{\alpha}_\mathbf{f}$ is known then we are free to use the full non-causal representation. We will also make the approximation that the determinant is independent of $\boldsymbol{\alpha}_\mathbf{f}$ in order to use the full non-causal model; this appears not to give too many problems in practice if we have a suitable prior for $\boldsymbol{\alpha}_\mathbf{f}$ to ensure the parameters are stable.

6.5 Hyperprior Models

The image prior model that has been described depends upon three hyperparameters: $\boldsymbol{\alpha}_\mathbf{f}$, σ_ν, $\mathbf{m}_\mathbf{f}$. Because in the hierarchical Bayesian approach it is considered that these parameters are also unknowns, the next stage is specification of hyperpriors defining the distributions of their likely values. This helps to remove dependance on exact specification of the parameters in the image prior, allowing us to set a range of possible values according to our expert knowledge, with a certain degree of confidence. We use conjugate priors, which helps to ensure tractability when either marginalising or using VB or Gibbs sampling methods, as the resulting conditional distributions are of the same known forms.

The AR parameters are assumed to be Gaussian distributed.[3] While an uninformative uniform distribution may be used reliably for the means (there are a large number of observed data points available for their estimation), in the complete model, we constrain them to be positive by using a vague (large variance) rectified Gaussian, denoted by $\left(\mathcal{N}^{+}\left(\theta \mid \cdot\right)\right.$. These hyperprior distributions are then:

$$p\left(\boldsymbol{a}_{f} \mid \delta_{a_{f}}\right) = \prod_{r \in \{1 \cdots R\}} \mathcal{N}\left(\boldsymbol{a}_{f_{r}} \,\middle|\, 0, \delta_{a_{f_{r}}} \mathbf{I}_{L_{a_{f}}}\right) \tag{6.43}$$

$$p\left(\boldsymbol{\mu}_{f} \mid \delta_{\mu_{f}}\right) = \prod_{r \in \{1 \cdots R\}} \mathcal{N}^{+}\left(\mu_{f_{r}} \mid 0, \delta_{\mu_{f}}\right) \tag{6.44}$$

For the hyperparameters which are variances of Gaussian distributions, the standard conjugate priors are IG distributions (see §4.1.6);[4] thus we have:

$$p\left(\boldsymbol{\sigma}_{v} \mid \boldsymbol{\alpha}_{v}, \boldsymbol{\beta}_{v}\right) = \prod_{r \in \{1 \cdots R\}} \mathcal{IG}\left(\sigma_{v_{r}}^{2} \mid \alpha_{v_{r}}, \beta_{v_{r}}\right) \tag{6.45}$$

$$p\left(\sigma_{w}^{2} \mid \alpha_{w}, \beta_{w}\right) = \mathcal{IG}\left(\sigma_{w}^{2} \mid \alpha_{w}, \beta_{w}\right) \tag{6.46}$$

6.5.1 Reparameterisation of Inverse-Gamma distribution

For convenience in specifying the priors for the IG distributions, we use a similar idea as in [153] to reparameterise the PDFs in terms of their expected value and a confidence parameter. This also gives a more intuitive meaning when updating the distributions in the VB approach.

We will see that the conditional distribution in the posterior resulting from multiplying two IGs, $\mathcal{IG}\left(\theta \mid \alpha_{\theta}, \beta_{\theta}\right)$ and $\mathcal{IG}\left(\theta \mid x, y\right)$, is proportional to $\mathcal{IG}\left(\theta \mid \alpha, \beta\right) = \mathcal{IG}\left(\theta \mid \alpha_{\theta} + x, \beta_{\theta} + y\right)$; α_{θ} and β_{θ} come from the prior, whose mean $\frac{\beta_{\theta}}{\alpha_{\theta}-1}$ we will now denote by $\mathbb{E}^{0}\left[\theta\right]$. Then to find the mean of the updated distribution, we apply the following identity for an IG distribution:

$$\mathbb{E}\left[\theta\right] = \frac{\beta}{\alpha - 1} = \frac{\beta_{\theta} + y}{\alpha_{\theta} + x - 1} \tag{6.47}$$

$$= \underbrace{\left(\frac{\alpha_{\theta} - 1}{\alpha_{\theta} + x - 1}\right)}_{\gamma_{\theta}} \underbrace{\left(\frac{\beta_{\theta}}{\alpha_{\theta} - 1}\right)}_{\mathbb{E}^{0}[\theta]} + \underbrace{\left(\frac{x}{\alpha_{\theta} + x - 1}\right)}_{1-\gamma_{\theta}} \left(\frac{y}{x}\right) \tag{6.48}$$

[3] They should be constrained to lie within a stability region, but in practice it is usually sufficient to constrain their variance by setting the hyperparameters suitably to ensure stability of the model. Improved modelling of the AR parameters is a possibility for future work.

[4] we could equally model the inverse of the variances with Gamma distributions and obtain similar results

which is a combination of the prior mean and the term $\frac{y}{x}$ weighted by a confidence parameter, γ_θ, which takes values between zero and one. Thus we may supply $\mathbb{E}^0[\theta]$ and γ_θ in place of α_θ and β_θ. The reasons for this procedure will become more clear when we consider the VB approach.

6.6 Blur model

So far we have described the nonstationary model for image representation. However, with appropriate selection of the hyperparameters, this model can also be used as a very flexible general blur prior

For instance, it includes as a special simplified case the stationary SAR model used in [153], where only a global variance representing PSF smoothness is estimated. Setting the hyperparemeters to encourage a larger amount of smoothing can efficiently represent a Gaussian or other types of blurs such as atmospheric turbulence or softening due to spherical aberration, which typically do not possess sharp features.

For other types of blurs, such as out-of-focus blur where there are sharp edges due to the shape of the lens aperture, a non-stationary blur with means and AR parameters can effectively represent these features.

Although we will consider subsets of the model, the full blur model is defined analogously to the image model; we partition the PSF support $\mathcal{S}_h$ into R_h blocks, each denoted $\mathcal{R}_{h_r}$. Then the model is defined, where all parameters have similar meanings as their counterparts in the image model, as:

$$p(\mathbf{h} \mid \mathbf{a}_h, \sigma_u, \mu_h) = \mathcal{N}^+\left(\mathbf{h} \,\middle|\, \bar{\mathbf{h}}, \boldsymbol{\Sigma}_h\right), \tag{6.49}$$

where
$$\boldsymbol{\Sigma}_h = \mathbb{E}\left[(\mathbf{h}-\bar{\mathbf{h}})(\mathbf{h}-\bar{\mathbf{h}})^T\right] = \mathbf{C}_h^{-1}\mathbf{Q}_u\mathbf{C}_h^{-T} \tag{6.50}$$

and
$$\bar{\mathbf{h}} = \mathbf{C}_h^{-1}\boldsymbol{\eta}_h, \tag{6.51}$$

$$p\left(\mathbf{a}_{h_r} \mid \delta_{a_{h_r}}\right) = \mathcal{N}\left(\mathbf{a}_{h_r} \,\middle|\, 0, \delta_{a_{h_r}}\mathbf{I}_{L_{a_h}}\right), \tag{6.52}$$

$$p\left(\mu_{h_r} \mid \delta_{\mu_h}\right) = \mathcal{N}^+\left(\mu_{h_r} \mid 0, \delta_{\mu_h}\right), \tag{6.53}$$

$$p\left(\sigma_{u_r}^2 \mid \alpha_{u_r}, \beta_{u_r}\right) = \mathcal{IG}\left(\sigma_{u_r}^2 \mid \alpha_{u_r}, \beta_{u_r}\right), \tag{6.54}$$

In practice for the initial experiments we use a much simpler blur model. In the MMAP experiments in Chapter 7, for instance, the prior is uniform, except we constrain the parameter space to a more restricted set by estimating the inverse of the blur. In the subsequent experiments, we begin with a classical smoothness prior using the discrete Laplacian, i.e. fixing the matrix $\mathbf{C}_{\mathrm{h}}$ and only estimating a single variance. The full model will become useful when considering estimation of all the parameters with Gibbs sampling in Chapter 10.

6.6.1 Positivity constraints

One other important physical phenomenon that is commonly imposed on the blur is the constraint of positivity of the PSF values, due to energy conservation. We will also make use of this constraint in Chapter 10. This constraint greatly reduces the ambiguity in the solution space, although is difficult to impose correctly as part of the model. In the past, this constraint has been imposed deterministically, for instance as part of a POCS type algorithm in [117]. Here we propose actually modelling the blur prior PDF as a distribution that only admits positive values; although we could investigate other distributions such as the log-normal, we will consider the use of a rectified normal distribution.

6.7 Complete Model

The full model that will be used in Chapter 10 is shown in the directed acyclic graph (DAG) in fig. 6.8. The various experiments in the remaining chapters used different simplifications of this complete model.

6.8 Conclusions

In this chapter, the new image and blur prior models that will be used in the coming chapters in this thesis have been presented. We began by discussing some of the features and problems of existing models, and the notions of why nonstationarity and higher order modelling is useful. We then reviewed existing stationary AR models and then described in a generative sense the new nonstationary model. Applying this concept in reverse, we consider the *excitation* signal that is responsible for generating the image.

We may fit the model to a real image and estimate this excitation, and hope that it is a good fit

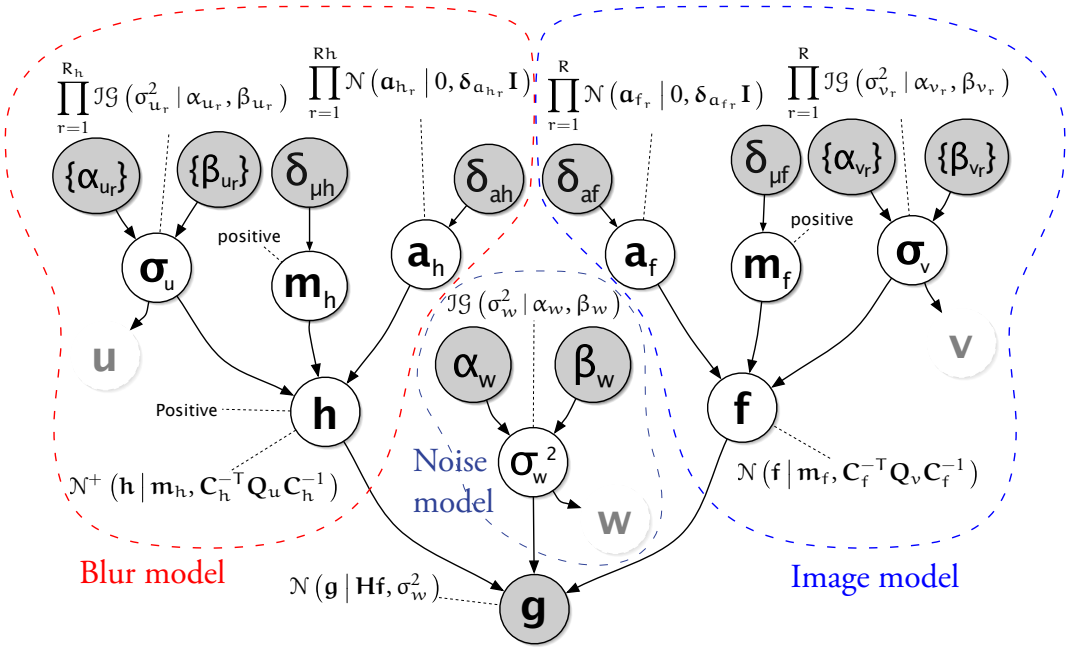


Figure 6.8: *Graphical model showing relationships between variables*

to the distribution prescribed by the model. Many image models take the approach of trying to fit a *heavy tailed* distribution to model the non-Gaussianity of the excitation, after filtering with some type of differencing operator. We show our model performs a similar goal in the opposite manner, by finding filters that try to Gaussianise the signal. Moreover, because we are using filters with a larger support than local difference operators, we stand a better chance of reducing correlation in the residual, making it white as is expected from the model, whereas methods that only model the heavy-tailed nature of the differences do not take the residual correlation into account. This is one explanation for why we can better model texture in the image.

In the remainder of the chapter, we described the probabilistic formulation of the model, beginning with a case when we do not include local means, then the case with the means. We also described the hyperprior distributions which are used in the hierarchical Bayesian framework to model the unknown hyperparameters in the model. Finally we described that the image model may also work as a reasonable blur prior in certain cases, even though it was designed to model natural images.

Chapter 7
MMAP method

In this chapter, Bayesian marginalisation techniques are employed in order to estimate the blur parameters directly from the degraded image. This extends the methods in [93] to two dimensions. However, certain approximations are needed in order to make this analysis possible; this means that the resulting algorithm is not so suitable for application to real images. Nevertheless, it provides some useful insight into the BID problem, and gives a foundation for the methods that will be investigated in the remaining chapters. The method described in this chapter was presented in [24].

One of the standard approaches to performing parameter estimation is MAP estimation (see §5.2.1). For the BID problem, direct maximisation of the posterior distribution is intractable due to dimensionality and high non-linearity. Hence alternating minimization (AM) is usually applied to linearise the optimisation and obtain an approximate solution. The Bayesian methodology lets us reduce the parameter space through marginalisation, resulting in the maximum marginalised *a posteriori* (MMAP) method: optimisation is now performed over the cost function defined on the smaller parameter space.

The principle of the MMAP method is to develop an image model that facilitates integrating out the nuisance parameters — parameters in the model whose actual values are not really of interest. Although we could, for instance, estimate the AR parameters in the model from the degraded image, then use these to restore the image in a two-step procedure, this would give a sub-optimal result, relying on a single estimate of these parameters. The process of marginalisation analytically takes care of accounting for the range of possible values these parameters may have when performing any further estimates, without the need to explicitly know their values. The only problem is that the models must be chosen very carefully in order to allow the required integrals to be found analytically, or else we must rely on numerical techniques such as Markov chain Monte Carlo (MCMC) (see Chapter 10).

The model that will be used in this chapter will allow for marginalisation of the image, and the image model parameters themselves, resulting in a PDF for the blur parameters (here denoted

as $\mathbf{h}'$) given the observed image, $p(\mathbf{h}' \mid \mathbf{g})$. This is, in a sense, in the category of *a-priori* blur estimation methods, because this blur estimate will then be used to restore the image $\mathbf{f}$, which was itself marginalised at the start of the process.

The blur model that is used here has a different parameterisation from the rest of the thesis. This parameterisation in fact model the inverse of the blur, so that effectively a restoration filter is being estimated. Consequently these parameters — which will be denoted $\mathbf{h}'$ in order to avoid confusion with the regular PSF coefficients $\mathbf{h}$ — lie in the same parameter space as the AR image model parameters. This concept will be used to illustrate why the nonstationary model can help in estimating the blur parameters. These ideas are due to the related method for audio used in [93], where it is natural to assume AR source and channel models.

As there is a deterministic relationship between $\mathbf{h}$ and $\mathbf{h}'$, we will continue to write the distributions conditional on $\mathbf{h}$ and transform to the parameter space of $\mathbf{h}'$ to perform the final optimisation. In the following sections, firstly the posterior distribution for the model used in this chapter is found, followed by a description of the marginalisation, and then a description of how the optimisation is performed in the parameter space of $\mathbf{h}'$. Finally examples of blur estimation for several synthetically generated images are presented, demonstrating the performance of the method; these show how the nonstationary image model affects the blur estimation results.

7.1 System model and MMAP method

In this chapter, a slightly simplified hierarchical model is used compared to the complete model; in this section the distributions and model approximations necessary to the find the resulting posterior are developed. Firstly the true image is marginalised from the model in §7.1.1, before discussing an approximation in §7.1.2 necessary to find a posterior amenable to marginalisation of the nuisance parameters, which is found in §7.1.4.

7.1.1 Marginal Likelihood

To begin the formulation, a reinterpretation of the likelihood used for ML parameter estimation in §6.1.1 is provided. This likelihood, $p\left(\mathbf{g} \mid \boldsymbol{\alpha}_f, \mathbf{h}, \sigma_v, \sigma_w^2\right)$, of equation (6.5), was found by a probability transformation of the prior for $\mathbf{f}$, and the addition of the noise as the

combination of two Gaussian random variables. This effectively eliminated $\mathbf{f}$ from the model (as shown in fig. 7.1), and therefore this likelihood may be viewed as the *marginal* likelihood. That is, the complete likelihood includes $\mathbf{g}$ and $\mathbf{f}$. In this section, an alternative derivation for the marginal likelihood is shown.

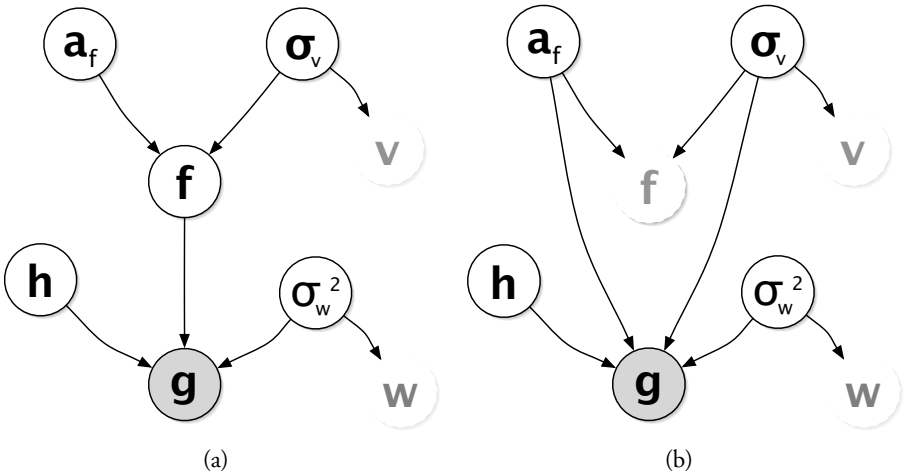


Figure 7.1: *(a) Initial observation & prior image model for the MMAP method; (b) After marginalising the image* $\mathbf{f}$.

Consider the conditional likelihood obtained from the standard observation model (4.3):

$$\mathrm{p}\left(\mathbf{g}\,|\,\mathbf{f},\mathbf{h},\sigma_v\right)=\mathcal{N}\left(\mathbf{g}\,|\,\mathbf{H}\mathbf{f},\sigma_v\right)=\sqrt{\frac{\det|\mathbf{Q}_w|^{-1}}{(2\pi)^{\mathrm{L_g}}}}\exp\Big[-\frac{1}{2}\|\mathbf{g}-\mathbf{H}\mathbf{f}\|^2_{\mathbf{Q}_w^{-1}}\Big] \tag{7.1}$$

By applying the AR model prior distribution for $\mathbf{f}$ from (6.2),

$$\mathrm{p}\left(\mathbf{f}\,|\,\boldsymbol{\alpha}_\mathrm{f},\sigma_v\right)=\mathcal{N}\left(\mathbf{f}\,|\,\mathbf{0},\boldsymbol{\Sigma}_\mathrm{f}\right)=\sqrt{\frac{\det|\mathbf{I}-\mathbf{A}|^2}{(2\pi)^{\mathrm{L_+}}\det|\mathbf{Q}_v|}}\exp\left[-\frac{1}{2}\mathbf{f}^{\mathsf{T}}(\mathbf{I}-\mathbf{A})^{\mathsf{T}}\mathbf{Q}_v^{-1}(\mathbf{I}-\mathbf{A})\mathbf{f}\right], \tag{7.2}$$

we obtain the following joint distribution (or complete likelihood, using the EM terminology), represented in the graphical model of fig. 7.1(a):

$$\mathrm{p}\left(\mathbf{g},\mathbf{f}\,\middle|\,\boldsymbol{\alpha}_\mathrm{f},\mathbf{h},\sigma_v,\sigma_w^2\right)=\mathrm{p}\left(\mathbf{g}\,\middle|\,\mathbf{f},\mathbf{h},\sigma_w^2\right)\mathrm{p}\left(\mathbf{f}\,|\,\boldsymbol{\alpha}_\mathrm{f},\sigma_v\right) \tag{7.3}$$

$$=\sqrt{\frac{\det|\mathbf{I}-\mathbf{A}|^2}{(2\pi)^{\mathrm{L_g}+\mathrm{L_+}}\det|\mathbf{Q}_v\mathbf{Q}_w|}}\exp\Big[-\frac{1}{2}\Big((\mathbf{f}^{\mathsf{T}}\mathbf{H}^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{H}\mathbf{f}-2\mathbf{g}^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{H}\mathbf{f}+\mathbf{g}^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{g})+(\mathbf{f}^{\mathsf{T}}\Sigma_\mathrm{f}^{-1}\mathbf{f})\Big)\Big]. \tag{7.4}$$

We will make use of the following Gaussian integral result: for an arbitrary matrix $\boldsymbol{\Gamma}$, vector $\boldsymbol{\beta}$ and scalar α, based on the normalisation of the Gaussian distribution, we have that [80]:

$$\begin{aligned}&\int_{\mathbb{R}^{P\times 1}}\cdots\int \exp\left[-\frac{1}{2}\left(\mathbf{x}^{\mathsf{T}}\boldsymbol{\Gamma}\mathbf{x}-2\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}+\alpha\right)\right]\mathrm{d}\mathbf{x}\\&=\frac{(2\pi)^{P/2}}{\sqrt{\det|\boldsymbol{\Gamma}|}}\exp\left[-\frac{1}{2}\left(\alpha-\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\Gamma}^{-1}\boldsymbol{\beta}\right)\right],\end{aligned} \tag{7.5}$$

Using this result, with $\boldsymbol{\Gamma}=\boldsymbol{\Sigma}_{\mathbf{f}}^{-1}+\mathbf{H}^{\mathsf{T}}\mathbf{Q}_{w}^{-1}\mathbf{H}$, $\boldsymbol{\beta}=\mathbf{H}^{\mathsf{T}}\mathbf{Q}_{w}^{-1}\mathbf{g}$ and $\alpha=\mathbf{g}^{\mathsf{T}}\mathbf{Q}_{w}^{-1}\mathbf{g}$, $\mathbf{f}$ can be marginalised from Equation (7.3), to give the same result as in (6.5), the marginal likelihood:

$$\mathrm{p}\left(\mathbf{g}\,\middle|\,\boldsymbol{\alpha}_{\mathrm{f}},\boldsymbol{\hbar},\boldsymbol{\sigma}_{v},\sigma_{w}^{2}\right)=\int\cdots\int\mathrm{p}\left(\mathbf{g},\mathbf{f}\,|\,\boldsymbol{\alpha}_{\mathrm{f}},\boldsymbol{\hbar},\boldsymbol{\sigma}_{v},\sigma_{w}\right)\mathrm{d}\mathbf{f} \tag{7.6}$$

$$=\frac{1}{\sqrt{(2\pi)^{L_{\mathbf{g}}}\det|\mathbf{H}\boldsymbol{\Sigma}_{\mathrm{f}}\mathbf{H}^{\mathsf{T}}+\mathbf{Q}_{w}|}}\exp\left[-\frac{1}{2}\left(\mathbf{g}^{\mathsf{T}}\left\{\mathbf{Q}_{w}^{-1}-\mathbf{Q}_{w}^{-1}\left(\mathbf{H}^{-\mathsf{T}}\boldsymbol{\Sigma}_{\mathrm{f}}^{-1}\mathbf{H}^{-1}+\mathbf{Q}_{w}^{-1}\right)^{-1}\mathbf{Q}_{w}^{-1}\right\}\mathbf{g}\right)\right] \tag{7.7}$$

$$=\frac{1}{\sqrt{(2\pi)^{L_{\mathbf{g}}}\det|\mathbf{H}\boldsymbol{\Sigma}_{\mathrm{f}}\mathbf{H}^{\mathsf{T}}+\mathbf{Q}_{w}|}}\exp\left[-\frac{1}{2}\mathbf{g}^{\mathsf{T}}(\mathbf{H}\boldsymbol{\Sigma}_{\mathrm{f}}\mathbf{H}^{\mathsf{T}}+\mathbf{Q}_{w})^{-1}\mathbf{g}\right] \tag{7.8}$$

The final result here, (7.8), is found by applying the *matrix inversion lemma*.[1] This marginalisation process is represented in the graphical model of fig. 7.1(b).

7.1.2 Effect of additive noise — approximate ML model

As described in §6.1, the standard likelihood expression (7.8) is highly non-linear in the parameters, so direct maximisation is prohibitive; furthermore it is not in a form where it is possible to analytically to marginalise $\boldsymbol{\alpha}_{\mathrm{f}}$ and $\boldsymbol{\sigma}_{v}$, due to the presence of $\mathbf{Q}_{w}$, which depends on σ_{w}^{2}. Therefore for the MMAP method, the approximation that the additive noise $\boldsymbol{w}$ is negligible will be assumed, i.e. $\sigma_{w}^{2}\to 0$ such that $\mathbf{Q}_{w}$ can be omitted.

Firstly, dropping the noise covariance $\mathbf{Q}_{w}$ from (7.8), the determinants become separable:

$$\mathrm{p}\left(\mathbf{g}\,|\,\boldsymbol{\alpha}_{\mathrm{f}},\boldsymbol{\hbar},\boldsymbol{\sigma}_{v}\right)=\frac{\det|\mathbf{I}-\mathbf{A}|}{\det|\mathbf{H}|\sqrt{(2\pi)^{L_{\mathbf{g}}}\det|\mathbf{Q}_{v}|}}\exp\left[-\frac{1}{2}\mathbf{g}^{\mathsf{T}}(\mathbf{H}\boldsymbol{\Sigma}_{\mathrm{f}}\mathbf{H}^{\mathsf{T}})^{-1}\mathbf{g}\right]. \tag{7.10}$$

[1] This Lemma is stated for some arbitrary matrix $\mathbf{M}$ and a positive definite matrix $\mathbf{X}\mathbf{X}^{\mathsf{T}}$ as:

$$(\mathbf{M}^{-1}+\mathbf{X}\mathbf{X}^{\mathsf{T}})^{-1}=\mathbf{M}-\left(\mathbf{M}\mathbf{X}(\mathbf{I}+\mathbf{X}^{\mathsf{T}}\mathbf{M}\mathbf{X})^{-1}\right)\mathbf{X}^{\mathsf{T}}\mathbf{M}=\mathbf{M}-\mathbf{M}\left((\mathbf{X}\mathbf{X}^{\mathsf{T}})^{-1}+\mathbf{M}\right)^{-1}\mathbf{M}. \tag{7.9}$$

The matrices $\mathbf{I} - \mathbf{A}$ and $\mathbf{H}$ will have a BTTB form in each case, which can make inference for these parameters significantly more challenging. One option is to approximate them with BCCB matrices (this is discussed for $\mathbf{H}$ in §7.4), however we do not do this but instead, a causal image model support is used in this chapter which further simplifies the expression. A causal support means the determinant will be lower triangular and equal to unity — otherwise it must be calculated in full. This gives a form of the likelihood leading to a posterior from which the AR parameters can be analytically marginalised:

$$p(\mathbf{g} \mid \boldsymbol{\alpha}_f, \mathbf{h}, \sigma_v) = \frac{1}{\det|\mathbf{H}|\sqrt{(2\pi)^{mn}\det|\mathbf{Q}_v|}} \exp\left[-\frac{1}{2}\hat{\mathbf{f}}^T(\mathbf{I}-\mathbf{A})^T\mathbf{Q}_v^{-1}(\mathbf{I}-\mathbf{A})\hat{\mathbf{f}}\right], \quad (7.11)$$

where the term $\mathbf{H}^{-1}\mathbf{g}$ has been denoted by $\hat{\mathbf{f}}$ for clarity. Note that this is *not* a probability transformation. The data matrix corresponding to $\hat{\mathbf{f}}$ is denoted $\hat{\mathbf{F}}$. $\hat{\mathbf{f}}$ is effectively an inverse filtered image estimate given a particular $\mathbf{h}$; we could in practice heuristically replace this with a regularised estimate, once we have identified the PSF coefficients. However our current goal is investigating the estimation of PSF parameters.

7.1.3 Prior distributions

To complete the hierarchical model for the MMAP method,the priors on the unknown model parameters need to be specified. The choice of priors described below lead to the final hierarchical model shown in fig. 7.2.

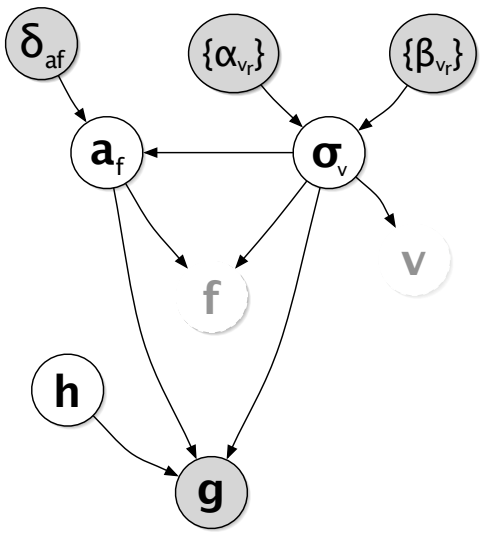


Figure 7.2: *Full hierarchical model for the MMAP method.*

7.1.3.1 Image variances

Two models are considered in this chapter for the image variances parameterising $\mathbf{Q}_v$. Firstly, the general case where the variances are different in each region, $\mathbf{Q}_{v'} = \text{diag}([\sigma_{v_1}^2 \mathbf{1}_{L_1}^T \cdots \sigma_{v_R}^2 \mathbf{1}_{L_R}^T]^T)$ as described in §6.3.4. The general prior for $\boldsymbol{\sigma}_v$ is the IG model described in §6.5, i.e.

$$p(\boldsymbol{\sigma}_v \,|\, \boldsymbol{\alpha}_v, \boldsymbol{\beta}_v) = \prod_{r\in\{1\cdots R\}} \mathcal{IG}\left(\sigma_{v_r}^2 \,|\, \alpha_{v_r}, \beta_{v_r}\right). \tag{7.12}$$

We assume here that the hyperparameters $\{\alpha_{v_r}, \beta_{v_r}\}$ are known and constant across the image, equal to $\{\alpha, \beta\}$. As $(\alpha, \beta) \to 0$ the distribution tends to σ^{-2}, an uninformative Jeffreys' prior. Although we can choose $\{\alpha, \beta\}$ to better reflect the likely range of variances, for simplicity at this stage the Jeffreys' Prior is used.

The second model is the simplified case where the variance $\sigma_{v_r}^2$ is constant across the whole image. In this case, (7.12) simply becomes

$$p(\boldsymbol{\sigma}_v \,|\, \alpha, \beta) = p\left(\sigma_v^2 \,|\, \alpha, \beta\right) = \mathcal{IG}\left(\sigma_v^2 \,|\, \alpha, \beta\right). \tag{7.13}$$

7.1.3.2 Image AR coefficients

The AR parameters in the BSAR model should be chosen from the bounded set that lead to stable 2-D filters, however this is a difficult constraint to impose analytically as a PDF. Instead, we will use the standard zero-mean Gaussian prior with known variance δ_{a_f} to approximate the requirement. However, to facilitate marginalisation of the variances $\boldsymbol{\sigma}_v$, we write this variance as a multiple of the local image variance $\sigma_{v_r}^2$ in each block, i.e. $\delta_{a_f}\sigma_{v_r}^2$ (this relation is indicated in fig. 7.2). This does not affect the degrees of freedom of the model and merely aids tractability. The resulting distribution is

$$p(\mathbf{a}_f \,|\, \delta_{a_f}) = \prod_{r\in\{1\cdots R\}} \mathcal{N}\left(\mathbf{a}_{f_r} \,\middle|\, 0, \delta_{a_f}\sigma_{v_r}^2 \mathbf{I}_{L_{a_f}}\right) \tag{7.14}$$

$$= \mathcal{N}(\mathbf{a}_f \,|\, \mathbf{0}, \delta_{a_f}\mathbf{R}_v) \tag{7.15}$$

where $\mathbf{R}_v$ is a diagonal matrix of size $L_{a_f} \times L_{a_f} = R \cdot L_{a_{f_r}} \times R \cdot L_{a_{f_r}}$, where L_{a_f} is the dimension of $\mathbf{a}_f$ and $L_{a_{f_r}}$ is the dimension of $\mathbf{a}_{f_r}$. In the single variance parameter case $\mathbf{R}_v = \sigma_v^2 \mathbf{I}_{R\cdot L_{a_f}}$, or in general, for uniform sized regions, $\mathbf{R}_v = \text{diag}(\boldsymbol{\sigma}_v) \otimes \mathbf{I}_{L_{a_f}}$.

In the limiting case, where the hyperparameter $\delta_{\mathbf{a}_f}$ tends to infinity, the distribution becomes uninformative. However to ensure stability of the resulting AR filters, using an empirically chosen value proves to be necessary. Although we do not show it here, it is also possible to use a hyperprior to model the distribution of the unknown $\delta_{\mathbf{a}_f}$ (see the discussion in §9.6).

7.1.3.3 Blur model

The issue of blur modelling has been discussed in Chapter 4. As a result, the prior for $\mathbf{h}$ depends on the particular blur model chosen. For simplicity in this chapter, however, we choose an uninformative uniform distribution over the coefficients $\mathbf{h}$ (and hence also over the actual estimation space of $\mathbf{h}'$). As will be seen in §7.1.7, we then constrain the blur estimates to a transformed and reduced parameter space defined by $\mathbf{h}'$.

7.1.4 Posterior PDF

Using the priors as described in §7.1.3, Bayes' Rule may be applied to the marginal likelihood to obtain the posterior pdf for the model parameters given the image estimate $\hat{\mathbf{f}}$ (which is implicitly calculated from the blurred image $\mathbf{g}$).

$$
\begin{aligned}
\mathrm{p}\left(\mathbf{h}, \mathbf{a}_f, \boldsymbol{\sigma}_v \,|\, \mathbf{g}\right) &= \frac{\mathrm{p}\left(\mathbf{g} \,|\, \mathbf{h}, \mathbf{a}_f, \boldsymbol{\sigma}_v\right) \mathrm{p}\left(\mathbf{h}, \mathbf{a}_f, \boldsymbol{\sigma}_v\right)}{\mathrm{p}\left(\mathbf{g}\right)} \\
&\propto \mathrm{p}\left(\mathbf{g} \,|\, \mathbf{h}, \mathbf{a}_f, \boldsymbol{\sigma}_v\right) \mathrm{p}\left(\mathbf{h}\right) \mathrm{p}\left(\mathbf{a}_f \,|\, \delta_{\mathbf{a}_f}, \boldsymbol{\sigma}_v\right) \mathrm{p}\left(\boldsymbol{\sigma}_v\right) \\
&\propto \frac{\det|\mathbf{Q}_v|^{-\frac{1}{2}} \det|\mathbf{R}_v|^{-\frac{1}{2}}}{(2\pi)^{\frac{(L_+ + L_{\mathbf{a}_f})}{2}} \det|\mathbf{H}|} \exp\left[-\frac{1}{2}(\hat{\mathbf{f}} - \hat{\mathbf{F}}\mathbf{a}_f)^{\mathsf{T}} \mathbf{Q}_v^{-1} (\hat{\mathbf{f}} - \hat{\mathbf{F}}\mathbf{a}_f) - \frac{\mathbf{a}_f^{\mathsf{T}} \mathbf{R}_v^{-1} \mathbf{a}_f}{2\delta_{\mathbf{a}_f}}\right] \mathrm{p}(\boldsymbol{\sigma}_v) \\
&\propto \frac{\mathrm{p}(\boldsymbol{\sigma}_v) \det|\mathbf{R}_v|^{-\frac{1}{2}}}{(2\pi)^{\frac{(L_+ + L_{\mathbf{a}_f})}{2}} \det|\mathbf{H}| \det|\mathbf{Q}_v|^{\frac{1}{2}}} \\
&\qquad \cdot \exp\left[-\frac{1}{2}\left(\mathbf{a}_f^{\mathsf{T}}\{\hat{\mathbf{F}}^{\mathsf{T}} \mathbf{Q}_v^{-1} \hat{\mathbf{F}} + \delta_{\mathbf{a}_f}^{-1} \mathbf{R}_v^{-1}\}\mathbf{a}_f - 2\mathbf{a}_f^{\mathsf{T}} \hat{\mathbf{F}}^{\mathsf{T}} \mathbf{Q}_v^{-1} \hat{\mathbf{f}} + \hat{\mathbf{f}}^{\mathsf{T}} \mathbf{Q}_v^{-1} \hat{\mathbf{f}}\right)\right]
\end{aligned}
\tag{7.16}
$$

Notice that since we only wish to find a function that can be maximised by varying the parameters, any constants of proportionality not dependant on the model parameters may be ignored. As there are two possible forms for $\mathrm{p}\left(\boldsymbol{\sigma}_v\right)$, we have not yet substituted for this variable, because they do not affect the first stage of marginalisation over the AR coefficients $\mathbf{a}_f$.

7.1.5 Marginalisation of nuisance parameters

Estimation of the AR model parameters and variances is not directly relevant to estimation of the blur parameters: they are regarded as *nuisance parameters* that may be removed by the process of *marginalisation* provided by the Bayesian framework. The principle is to partition the parameter vector into two sets — those of interest, and the nuisance parameters — and integrate the PDF w.r.t the nuisance parameters. In this case, the marginal PDF we desire is:

$$\mathrm{p}(\mathbf{h}|\mathbf{g}) = \int \cdots \int \mathrm{p}\left(\mathbf{h}, \boldsymbol{\alpha}_\mathrm{f}, \boldsymbol{\sigma}_\nu \,|\, \mathbf{g}\right) \mathrm{d}\boldsymbol{\alpha}_\mathrm{f} \cdot \mathrm{d}\boldsymbol{\sigma}_\nu \tag{7.17}$$

The integration is done in a similar manner to that described in [93], via standard Gaussian and Gamma integrals. Marginalisation under the two possible models for the excitation variances $\boldsymbol{\sigma}_\nu$ are considered in Appendix D, depending on whether a single variance is chosen for the whole image, or an individual variance parameter for each block. In the former case the result is:

$$\mathrm{p}\left(\mathbf{h} \,|\, \mathbf{g}\right) \propto \frac{1}{\det|\mathbf{H}|} \frac{\left(\hat{\mathbf{f}}^\mathsf{T}\hat{\mathbf{f}} - \hat{\mathbf{f}}^\mathsf{T}\hat{\mathbf{F}}(\hat{\mathbf{F}}^\mathsf{T}\hat{\mathbf{F}} + \delta_{\alpha_\mathrm{f}}^{-1}\mathbf{I}_{\mathrm{L}_{\alpha_\mathrm{f}}})^{-1}\hat{\mathbf{F}}^\mathsf{T}\hat{\mathbf{f}} + 2\beta\right)^{-\frac{\mathrm{L}_+ + 1 + 2\alpha}{2}}}{\det\left|\hat{\mathbf{F}}^\mathsf{T}\hat{\mathbf{F}} + \delta_{\alpha_\mathrm{f}}^{-1}\mathbf{I}_{\mathrm{L}_{\alpha_\mathrm{f}}}\right|^{\frac{1}{2}}}, \tag{7.18}$$

and in the latter,

$$\mathrm{p}(\mathbf{h}|\mathbf{g}) \propto \frac{1}{\det|\mathbf{H}|} \prod_{\mathrm{r}\in\{1\cdots\mathrm{R}\}} \left(\frac{\left(\hat{\mathbf{f}_\mathrm{r}}^\mathsf{T}\hat{\mathbf{f}_\mathrm{r}} - \hat{\mathbf{f}_\mathrm{r}}^\mathsf{T}\hat{\mathbf{F}}_\mathrm{r}(\hat{\mathbf{F}}_\mathrm{r}^\mathsf{T}\hat{\mathbf{F}}_\mathrm{r} + \delta_{\alpha_\mathrm{f}}^{-1}\mathbf{I}_{\mathrm{L}_{\alpha_{\mathrm{f}_\mathrm{r}}}})^{-1}\hat{\mathbf{F}}_\mathrm{r}^\mathsf{T}\hat{\mathbf{f}_\mathrm{r}} + 2\beta_{\nu_\mathrm{r}}\right)^{-\frac{\mathrm{L}_\mathrm{r} + 1 + 2\alpha\nu_\mathrm{r}}{2}}}{\det\left|\hat{\mathbf{F}}_\mathrm{r}^\mathsf{T}\hat{\mathbf{F}}_\mathrm{r} + \delta_{\alpha_\mathrm{f}}^{-1}\mathbf{I}_{\mathrm{L}_{\alpha_{\mathrm{f}_\mathrm{r}}}}\right|^{\frac{1}{2}}} \right). \tag{7.19}$$

7.1.6 Blur Parameter Estimation

To perform blur estimation with the MMAP method, the marginal posterior in (7.18) or (7.19) is maximised over the space of possible blurs. The complicated form of these equations means that numerical optimisation is required to solve this problem.

It is usual to take the negative log and minimise:

$$\hat{\mathbf{h}} = \arg\min_{\mathbf{h}} \left(-\log \mathrm{p}\left(\mathbf{h} \,|\, \mathbf{g}\right) \right). \tag{7.20}$$

For each blur $\hat{\mathbf{h}}$ under test, the procedure essentially consists of finding a restoration $\hat{\mathbf{f}}$ of the whole image using this particular $\hat{\mathbf{H}}$, then (7.19) effectively checks how well this conforms to our model. The term

$$\tilde{\mathbf{f}}_r = \hat{\mathbf{F}}_r \hat{\mathbf{a}}_r = \hat{\mathbf{F}}_r \left\{ \hat{\mathbf{F}}_r^T \hat{\mathbf{F}}_r + \delta^{-1} \mathbf{I}_{\frac{L_{a_f}}{L_+}} \right\}^{-1} \hat{\mathbf{F}}_r^T \hat{\mathbf{f}}_r \tag{7.21}$$

may be regarded as a prediction of $\hat{\mathbf{f}}_r$ using an estimate of the AR parameters for this block, $\hat{\mathbf{a}}_r$. Then the model fitting term given by $\hat{\mathbf{f}}^T(\hat{\mathbf{f}} - \tilde{\mathbf{f}})$, as used in the numerator of equation (7.19), is the prediction error weighted by the local image magnitude.

Using this notation, the optimisation procedure becomes in the single variance case[2]

$$\begin{aligned} \hat{\mathbf{h}} = \arg\min_{\mathbf{h}} \Bigg[& \frac{L_+ + 1 + 2\alpha}{2} \log \left\{ \sum_{r \in \{1 \cdots R\}} \left(\hat{\mathbf{f}}_r^T (\hat{\mathbf{f}}_r - \tilde{\mathbf{f}}_r) + 2\beta \right) \right\} \\ & + \log\det|\mathbf{H}| + \frac{1}{2} \sum_{r \in \{1 \cdots R\}} \log\det \left| \hat{\mathbf{F}}_r^T \hat{\mathbf{F}}_r + \delta_{a_f}^{-1} \mathbf{I}_{\frac{L_{a_f}}{R}} \right| \Bigg], \end{aligned} \tag{7.22}$$

and for the variance-per-block approach:

$$\begin{aligned} \hat{\mathbf{h}} = \arg\min_{\mathbf{h}} \Bigg[& \sum_{r \in \{1 \cdots R\}} \left\{ \frac{\frac{L_+}{R} + 1 + 2\alpha_{v_r}}{2} \log \left(\hat{\mathbf{f}}_r^T (\hat{\mathbf{f}}_r - \tilde{\mathbf{f}}_r) + 2\beta_{v_r} \right) \right\} \\ & + \log\det|\mathbf{H}| + \frac{1}{2} \sum_{r \in \{1 \cdots R\}} \log\det \left| \hat{\mathbf{F}}_r^T \hat{\mathbf{F}}_r + \delta_{a_f}^{-1} \mathbf{I}_{\frac{L_{a_f}}{R}} \right| \Bigg], \end{aligned} \tag{7.23}$$

7.1.7 Blur re-parameterisation: Implementation details

Although the method presented so far should apply to a general linear blur model, such as the commonly used non-causal MA PSF model, in the experiments in this chapter, a causal parametric AR blur is used. There are several reasons for this: firstly, choosing a causal model means that the complicated $\det|\mathbf{H}|$ term need not be evaluated; secondly, this conforms to work in [93] where the method has been extensively tested in the 1D case. Moreover, by letting the blur and image model share the same parameter space, we can observe some important features of the model and how nonstationarity can aid blur estimation.

Furthermore, due to the approximations made to derive the MMAP expressions, an inverse

[2] Note that we can decompose into the summation over blocks as shown here, because the matrix $\mathbf{F}$ is block diagonal), as described in Appendix D

filtering operation is used to find $\hat{\mathbf{f}} = \mathbf{H}^{-1}\mathbf{g}$. This does not incorporate any regularisation; however with an AR blur model, $\mathbf{H}^{-1}$ is well defined, and we are effectively estimating an MA restoration filter which does not severely amplify noise. A regularised estimate for $\hat{\mathbf{f}}$ could be used here, but as it is not part of the model, parameter estimates may not be optimal.

We will later briefly consider in §7.4 the MMAP approach with the standard non-causal (and MA) models, and describe the problems that are encountered. The methods presented in the rest of the thesis will largely avoid these problems with causality and also allow regularisation of the image.

Here we use a "deterministic" parametric approach whereby the set of parameters to be estimated in $\mathbf{h}$ directly generate the PSF according to the parametric AR model. As mentioned in the chapter introduction, we let $\mathbf{h}$ define the PSF used in $\mathbf{H}$ in the usual way, then let:

$$\mathbf{H} = (\mathbf{I} - \mathbf{H}')^{-1}, \tag{7.24}$$

where $\mathbf{H}'$ is another BTTB matrix parameterised by $\mathbf{h}'$. The estimation in (7.23) is then performed over the space of $\mathbf{h}'$. For the experiments presented here we constrain $\mathbf{h}'$ to a first-order causal case, i.e. $\mathbf{h}' = [h_1, h_2, h_3]^T$, which parameterises the defining kernel h' used to construct $\mathbf{H}'$ as

$$h'(k,l) = \begin{bmatrix} h_3 & h_2 \\ h_1 & \bullet \end{bmatrix}, \tag{7.25}$$

where the $\bullet$ marks the centre of the kernel image, which is always zero.

7.2 Experimental Results

A number of experiments that were performed using the MMAP method will now be described, along with conclusions reached. We begin by generating synthetic datasets according to the model chosen, and blurring with known PSFs, in order to validate and test the algorithms' performance, before attempting to proceed to cases where the data does not conform exactly to the model, and using real images. Results are shown for the selection of different block sizes and model orders.

7.2.1 Experiment 1. — AR-AR model, Single blocks

Firstly we consider the case of a synthetically generated image, with various numbers of blocks according to the BSAR model. We begin by using the AR-AR model (AR blur and AR source image) similar to that used in the blind deconvolution method for acoustic signals in [93].

To begin with, a single block is used, corresponding to a spatially invariant image $\mathbf{f}$. We consider a simple first order image model, with causal support, which matches that for the blur. That is, we let $\mathbf{a}_f = [a_1, a_2, a_3]^T$ and use the following defining kernel to generate $\mathbf{A}$:

$$a(k,l) = \begin{bmatrix} a_3 & a_2 \\ a_1 & \bullet \end{bmatrix}, \tag{7.26}$$

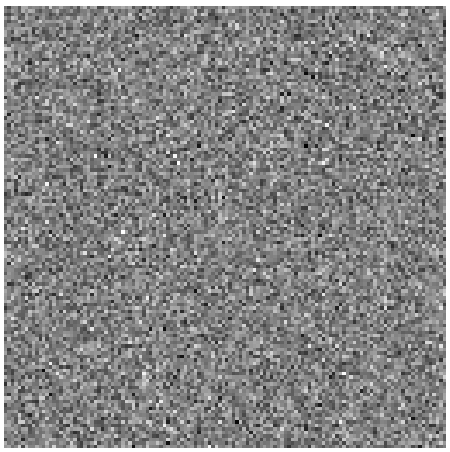
An example of a synthetic image $\mathbf{f} = (\mathbf{I} - \mathbf{A})^{-1}\boldsymbol{\nu}$ generated with this model, using is shown in fig. 7.3(c). The WGN realisation used to generate this, $\boldsymbol{\nu}$, is shown in fig. 7.3(a). The parameters for the source image were chosen as $\mathbf{a}^T = [0.45, 0.35, 0.025]$. When $\mathbf{f}$ is blurred according to $\mathbf{g} = (\mathbf{I} - \mathbf{H}')^{-1}\mathbf{f}$, the image in Fig. 7.3(d) is obtained. The parameters for the blur were chosen as $\mathbf{h}' = [0.55, -0.1, 0.025]^T$. The additive observation noise is not included at this stage. We now attempt to estimate the parameters of the blur.

Since there does not exist an equivalent to the fundamental theorem of algebra in dimensions higher than one [116], it is not generally possible to produce a factorisation of the polynomial space to produce a pole-zero type plot as may be done in 1D. Thus it is not easy to visualise on a 2D plot more than two parameters at once; hence we choose to assume one of the parameters, h_3 is known, and for these experiments only estimate h_1 and h_2. A contour plot in (h_1, h_2) space may be produced, showing the value we seek to minimise: the $-\log(\text{PDF})$ of the blur parameters, from (7.23). This is plotted over a grid of points centred around the (known) true value of the parameters[3]. Since the excitation signal $\boldsymbol{\nu}$ is random, each run of the simulation generates different realisations of the process.

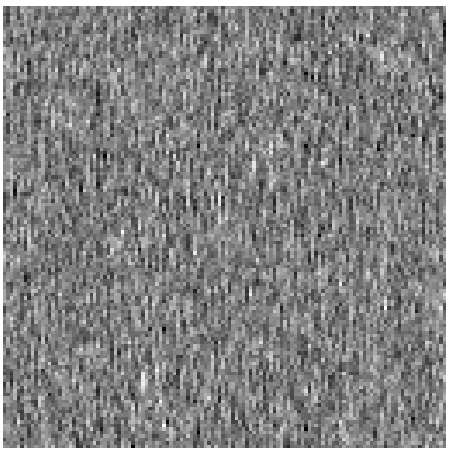
Fig. 7.4(a) shows $p(h_1, h_2)$ from a realisation where the minimum (marked with a circle) occurred at the wrong place, near to the source AR parameters (a_1, a_2) in fact.[4] When the

[3] In the plots shown, the $-\log(\text{PDF})$ has been mapped to the range 0–1 and compressed with a monotonic power law mapping $z \to z^\zeta, 0 < \zeta < 1$ to maintain a reasonable spacing of contours in flatter areas of interest near the minima (the contours may appear artificially close in the immediate proximity of the minima, as in fact the PDFs are quite flat here).

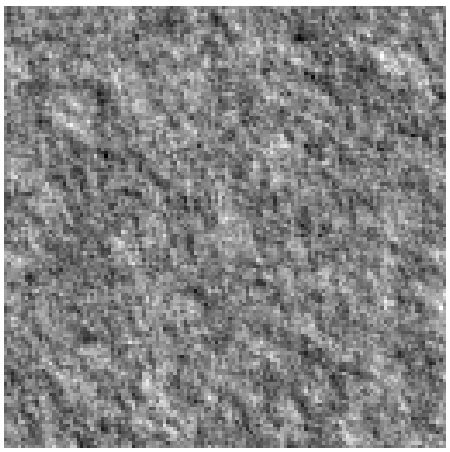
[4] It is important to realise that (h_1, h_2) share the same parameter space as (a_1, a_2), since we have chosen $h_3 = a_3$, and the two models share the same form. (a_1, a_2) is marked in the figure with a triangle.



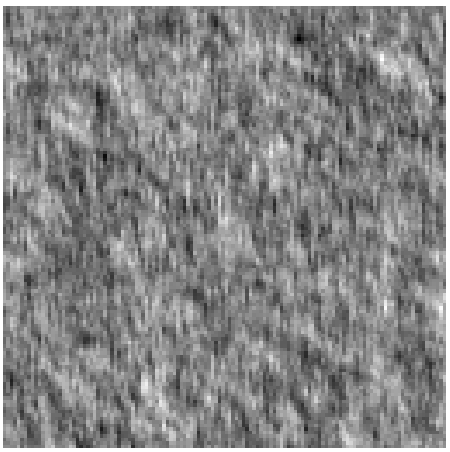
(a) Source excitation noise image, $\boldsymbol{\nu}$



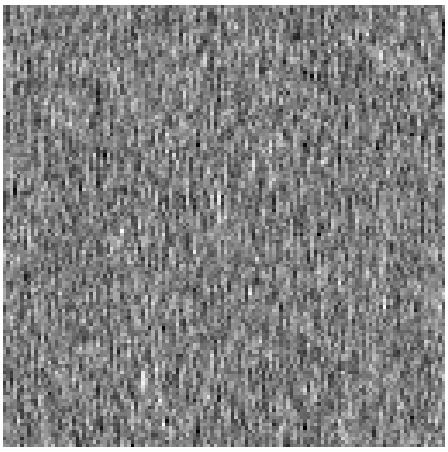
(b) $\boldsymbol{\nu}$ filtered with AR blur parameters



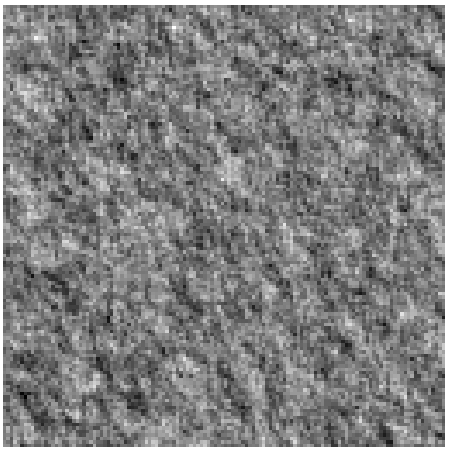
(c) Synthesised source image $\mathbf{f}$ ($\boldsymbol{\nu}$ filtered with source AR parameters)



(d) AR Blurred image $\mathbf{g}$ ($\mathbf{f}$ filtered with AR blur parameters)

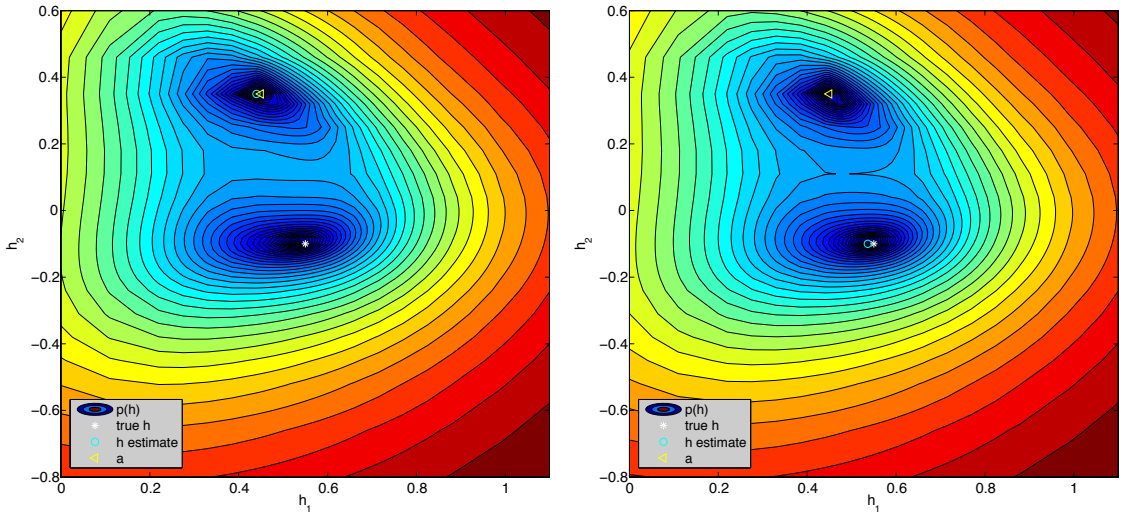


(e) $\mathbf{g}$ deblurred with incorrect parameters, corresponding to estimate of fig. 7.4(a)



(f) $\mathbf{g}$ deblurred with correct parameters, corresponding to estimate of fig. 7.4(b)

Figure 7.3: *Experiment 1. Single-block synthetic image,* 1^{st} *order AR source and blur*



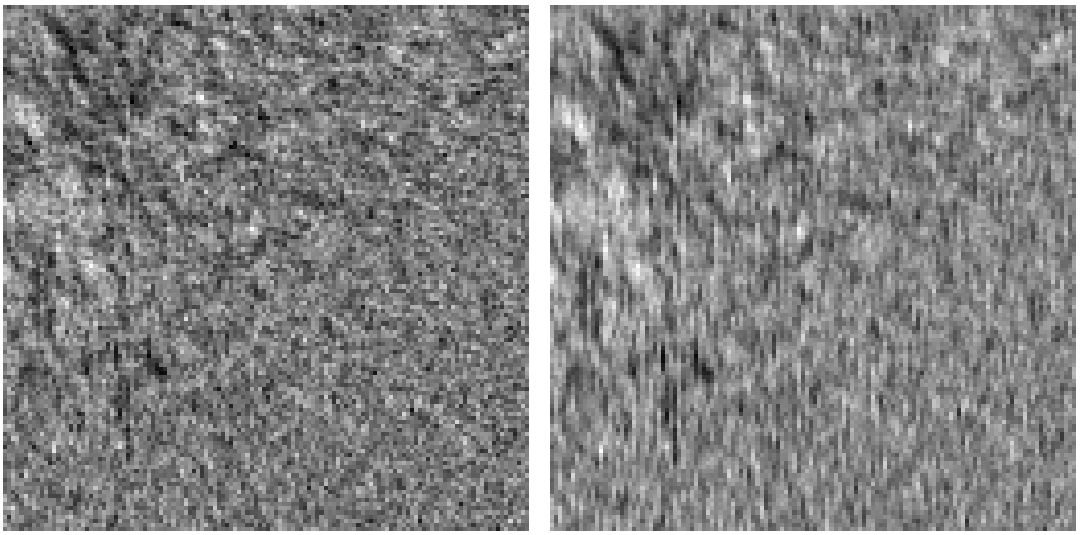
(a) Incorrectly estimated blur parameters (b) Correctly estimated blur parameters

Figure 7.4: *Experiment 1. Probability for AR blur parameters for single-block image. Contours show* $p(\mathbf{h}' \mid \mathbf{g})$*, with blue giving high probability and red low probability. The true value of* (h_1, h_2) *is marked with a white star, and* (a_1, a_2) *with a yellow triangle, while the estimate of* (h_1, h_2) *is indicated by a blue circle.*

blurred image is deblurred with these parameters, the image in fig. 7.3(e) is produced. Notice the similarity to the image being synthesised from $\boldsymbol{\nu}$ with the $\mathbf{a}_f$ parameters instead, as in fig. 7.3(b). In another realisation, the estimated $\mathbf{h}'$ parameters were near their true values (marked with a star), as shown in fig. 7.4(b), giving the deblurred image of fig. 7.3(f). These realisations occur with roughly equal probability since there is nothing to distinguish the two parameter sets when there is only a single block, i.e. there is an ambiguity between (a_1, a_2) and (h_1, h_2).

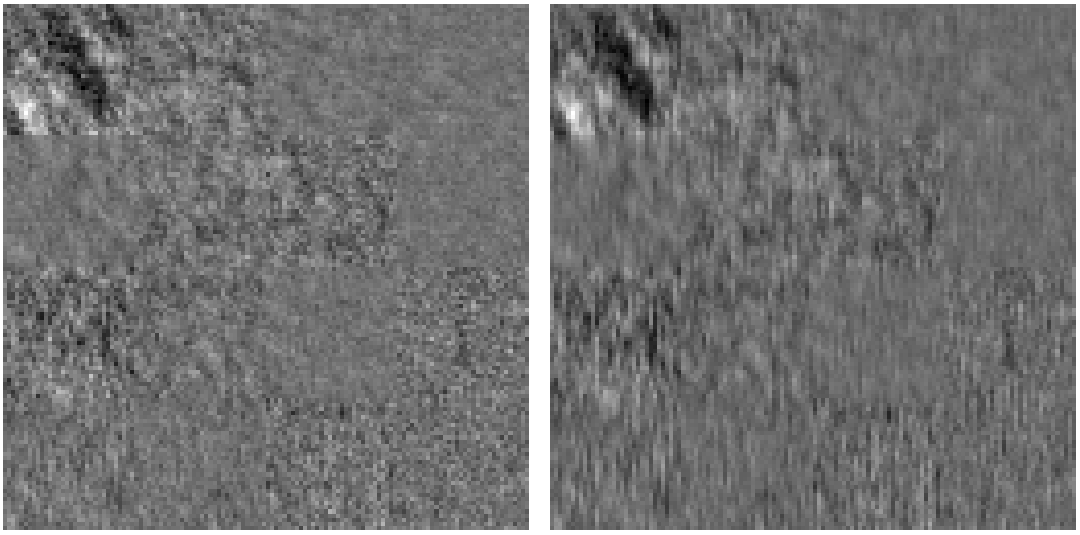
7.2.2 Experiment 2. — AR-AR model, Effect of multiple blocks

The experiment of §7.2.1 is repeated, but now different AR parameters for the source image are assigned to each of 16 different blocks (a 128×128 image is being used so each block is 32×32 pixels). These parameters are again marked as triangles on the probability plots (again they all share the same fixed value of a_3 as before). Firstly, the excitation variances are set the same in each block. This yields the source image of fig. 7.5(a), which is blurred with the same h as before to give fig. 7.5(b).



(a) **f** (b) **g**

Figure 7.5: *Experiment 2. 4×4-block image, same variance for each block*



(a) **f** (b) **g**

Figure 7.6: *Experiment 2. 4×4-block image, different variance for each block*

The using the two different formulations, assuming either an individual variance in each block, (7.23), or a global variance, (7.22) are now compared. These PDFs are evaluated and are shown in figs. 7.7(a) and 7.7(b) respectively. For these experiments, the results from 5 realisations have been averaged together to iron out fluctuations caused by using limited data sizes. As can be seen, there is very little difference between the two formulations, the correct model in this case (fig. 7.7(b)) having marginally tighter contours in the flat region around the minimum.

However when the source image actually does contain separate variances for each block, as in figs. 7.6(a) and 7.6(b), using the correct model of (7.23) gives a much superior result as can be seen from the plots of figs. 7.8(a) and 7.8(b). In general though, we can conclude that all of these estimates are more accurate than the case for a single source AR parameter set, or equivalently, the same parameters in each block.

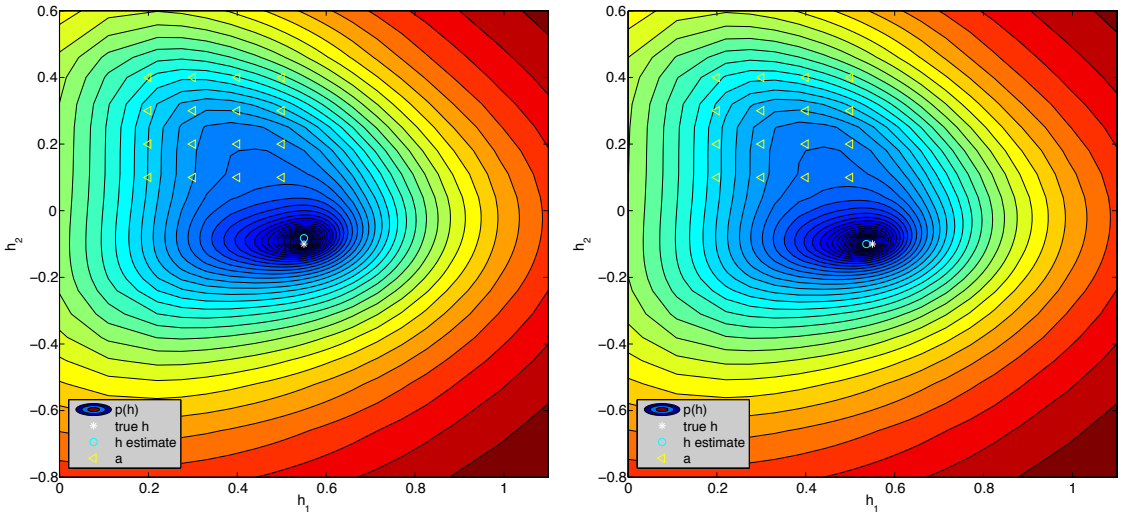
7.2.3 Experiment 3. — AR-AR model, Estimation using incorrect block size

If the block size in the source image is reduced, more local features can be modelled. However there is less data to use in each block to perform the estimation. An example using 256 blocks is shown in fig. 7.9(a), and the corresponding probability plot from the blurred image (not shown) in fig. 7.9(b). The variance-per-block formulation is used from now on. It can be seen that the estimate is correct, and the probability similar to the previous one, fig. 7.8(b), except the variance of the estimate is now much reduced due to the increased variability of the underlying source image.

When the incorrect block size is used, as in fig. 7.10(a), where the same 32×32 pixel block size as before is used to estimate the probability (although the source image is that of fig. 7.9(a)), it can be seen that the estimate is still almost correct, but there is a lot more uncertainty introduced.

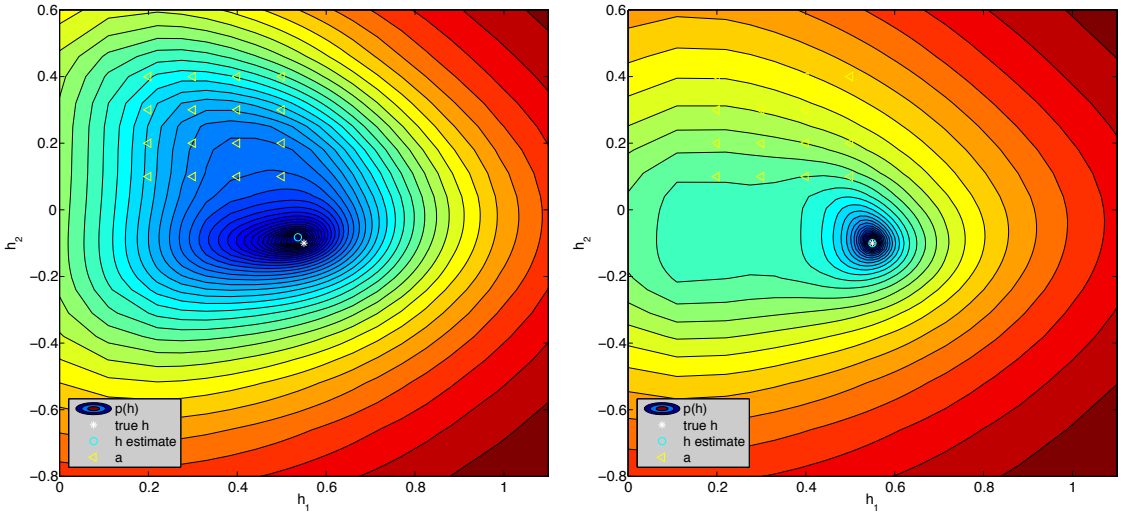
7.2.4 Experiment 4. — AR-AR model, Estimation using incorrect model order

Using a higher model order for the source parameters during estimation should again be able to represent more complicated images. Once more it would be expected that more data is required for a satisfactory estimate. In fig. 7.10(b), the previous experiment is repeated using the incorrect block size and also the wrong source AR model order (3 instead of 1). This



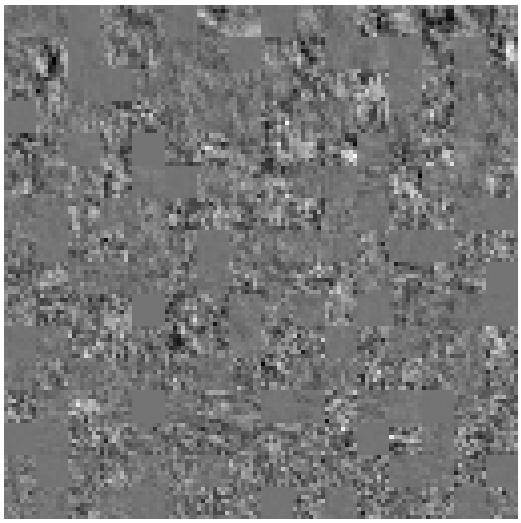
(a) $p(\mathbf{h}' \mid \mathbf{g})$ assuming individual variance per block (b) $p(\mathbf{h}' \mid \mathbf{g})$ assuming single variance for whole image

Figure 7.7: *Experiment 2. Probability for AR blur parameters for 4×4-block image, same variance for each block*

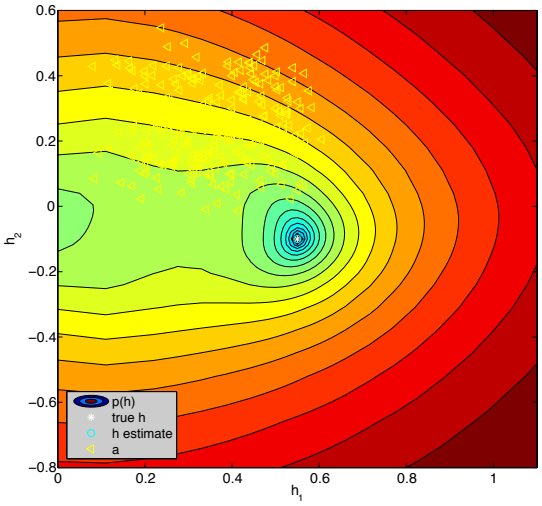


(a) $p(\mathbf{h}' \mid \mathbf{g})$ assuming single variance for whole image (b) $p(\mathbf{h}' \mid \mathbf{g})$ assuming individual variance per block

Figure 7.8: *Experiment 2. Probability for AR blur parameters for 4×4-block image, different variance for each block*

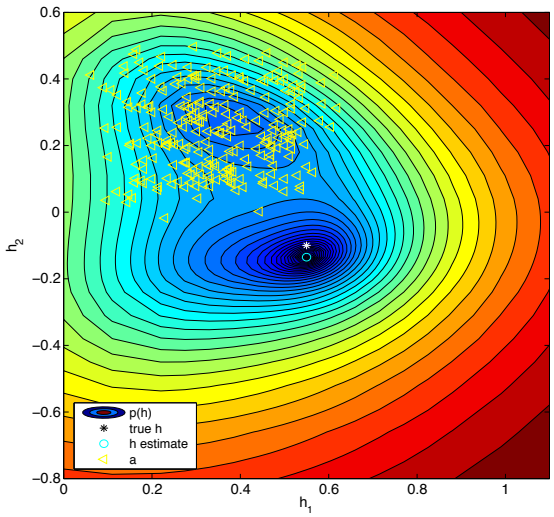


(a) $\mathbf{f}$, 8×8 pixel blocks

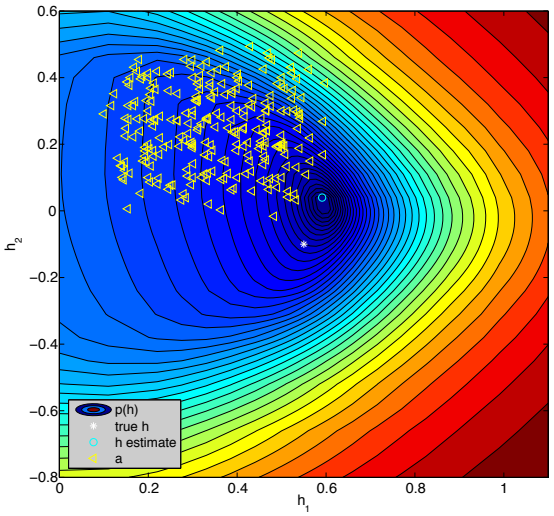


(b) $p(\mathbf{h}' \mid \mathbf{g})$ for fig. 7.9(a), estimated with correct block size (8×8 pixels) and model order (1)

Figure 7.9: *Experiment 3. 256-block (8×8 pixels) AR(1) image & probability for AR blur parameters*



(a) $p(\mathbf{h}' \mid \mathbf{g})$ for fig. 7.9(a), estimated with incorrect block size (32×32 pixels) but correct model order (1)



(b) $p(\mathbf{h}' \mid \mathbf{g})$ for fig. 7.9(a), estimated with incorrect model order (3) and block size (32×32 pixels)

Figure 7.10: *Experiment 3. Probability for AR blur parameters for 256-block image (8×8 pixel blocks)*

clearly shows unsatisfactory results.

When the correct block size but incorrect model order is used (it is overestimated as 3 again), the situation is improved, as in fig. 7.11(a). For comparison, if we generate a synthetic image $\mathbf{f}$ (not shown) with a source AR model order of 3, and estimate this correctly, the plot of fig. 7.11(b) is produced. The AR parameters cannot be plotted on this figure since they do not occupy the same parameter space as the blur parameters any more (they are drawn from a Gaussian distribution around $\mathbf{0}$). Also shown in fig. 7.12 are the results from the same experiment using 16 (32×32 pixel) blocks, with the correct assumptions about block size and model order; again we see that using larger blocks is not as effective as smaller blocks in general[5]. However we may also note that in general that as the AR model goes up, the total number of parameters to be estimated increases and the variance of the estimates increases. From this point of view, the best estimate was that of fig. 7.9(b), where the source and estimation orders are both 2. Unfortunately in conflict with this, we may need higher order models to better model real images.

7.2.5 Experiment 5. — AR-AR model, Effect of additive noise

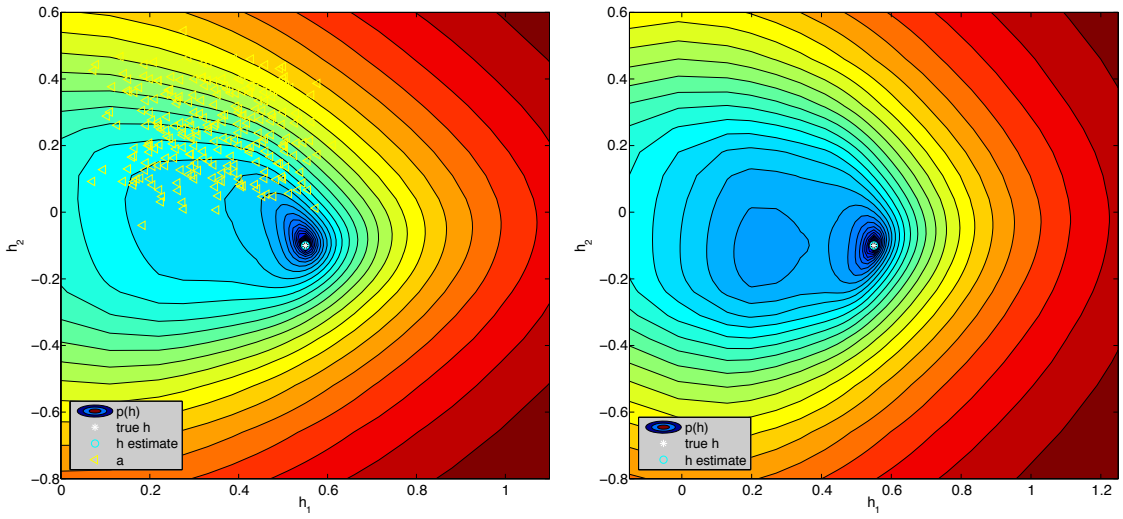
Recall that the models used in this chapter do not account for observation noise. The effect on the blur estimates of introducing the noise $\boldsymbol{w}$ back into the test image is now considered. Using the standard measure of blurred-image SNR (BSNR), defined as:

$$\text{BSNR} = 10 \log_{10} \frac{\text{Var}[\mathbf{H}\mathbf{f}]}{\text{Var}[\boldsymbol{w}]} \text{ (dB)}, \tag{7.27}$$

some of the previous experiments are repeated with additive WGN. In fig. 7.13(a), we see a 10dB BSNR image still yields a reasonable estimate in the case of a correctly chosen AR(1) model with 256 blocks (compare to fig. 7.9(b)). For 16 blocks and 10dB BSNR in fig. 7.13(b), the estimate is less certain (compare to fig. 7.8(b)). Down to 10dB, the variation for both was fairly minor. However at 5dB there is a sudden drop in performance, shown in fig. 7.14 for the same two cases, where it can be seen fig. 7.14(b) has already 'broken down', and fig. 7.14(a) is just about to do the same (at 0dB it has a similar appearance).

With the higher order model, it is found that this sudden drop does not seem to occur. Fig.

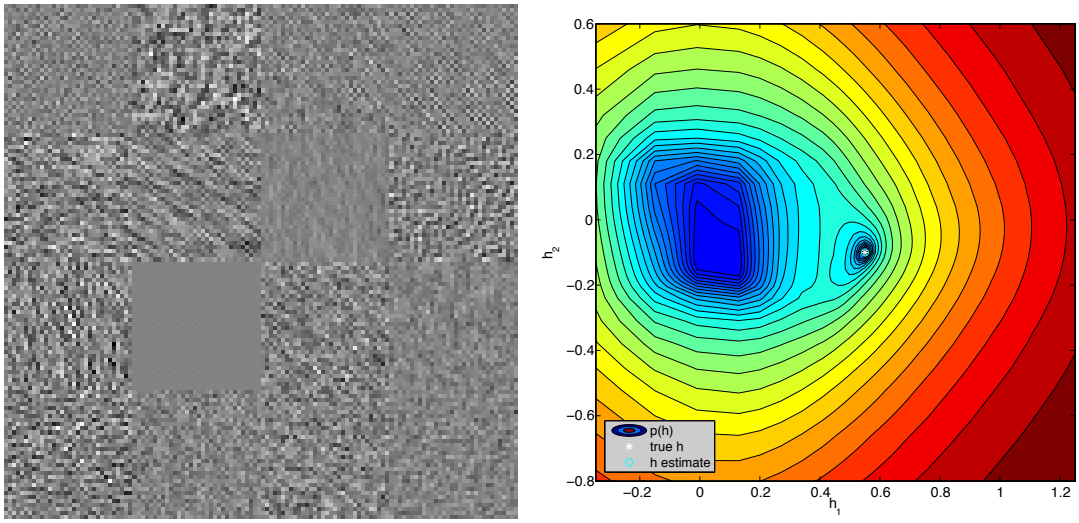
[5]Also observe here the cost function is multi-modal and a non-linear optimisation method would be required to find the global minimum.



(a) $p(\mathbf{h}' \mid \mathbf{g})$ for fig. 7.9(a), estimated with incorrect model order (3) but correct block size (8×8 pixels)

(b) $p(\mathbf{h}' \mid \mathbf{g})$ for AR(3) source, estimated with correct model order (3) and block size (8×8 pixels)

Figure 7.11: *Experiment 4. Probability for AR blur parameters for 256-block image (8×8 blocks)*



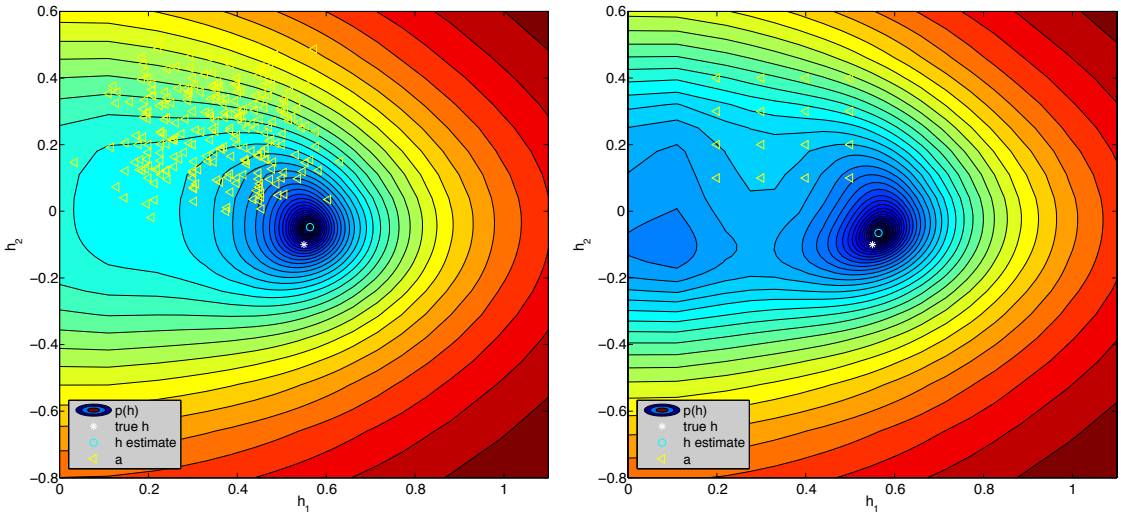
(a) $\mathbf{f}$, 32×32 pixel blocks, 3rd order model

(b) $p(\mathbf{h}' \mid \mathbf{g})$ for fig. 7.12(a), estimated with correct block size (32×32 pixels) and model order (3)

Figure 7.12: *Experiment 4. 16-block (32×32 pixels) AR(3) image & probability for AR blur parameters*

7.15(a) shows the experiment with correctly chosen AR(3) model and 16 blocks (compare to fig. 7.12(b)) but with 5dB BSNR. It is observed that as the BSNR decreases, the parameter estimate drifts towards zero. This may be attributed to the fact that the source image parameters no longer occupy the same parameter space as the blur parameters, and the higher order may mean that the more structured source image is more 'visible' in the noise.

Fig. 7.15(b) shows the experiment with 256 blocks and incorrectly chosen AR(3) model where the source order was 1 (compare to fig. 7.11(a)) but with 5dB BSNR. The results in this case seem fairly promising, with some degree of accuracy still being obtained at the 5dB level and the estimate drifting again towards $\mathbf{0}$ as the BSNR plunges below 0dB. Results for deblurring an image in this case with 10dB BSNR are shown in fig. 7.16. As can be seen, the image is reasonably accurately deblurred even though the estimate of the parameters was not entirely correct at $\hat{\mathbf{h}} = [0.585, 0.005, 0.025]$, and there is a moderate but not excessive amount of noise amplification over the blurred image visible. Below 5dB, the amplified noise dominates the visual restoration and it is hard to make out any features, even though the error in the parameter estimates is not at its worst. Overall it seems the parameter estimates are robust in the presence of noise[6], assuming they were accurate in the first place in the noiseless case.

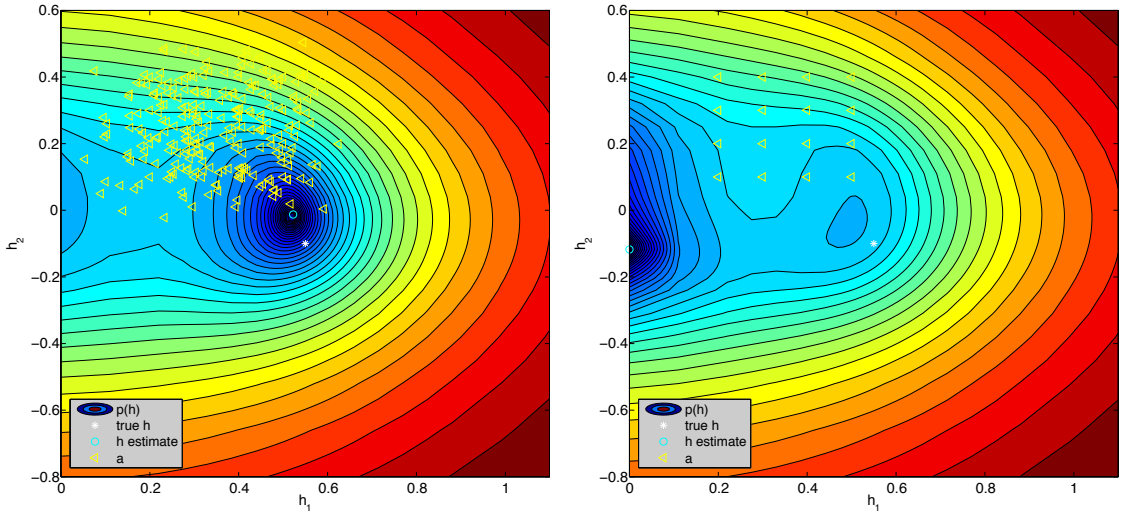


(a) as fig. 7.9(b) but with 10dB BSNR

(b) as fig. 7.8(b) but with 10dB BSNR

Figure 7.13: *Experiment 5. Effect of additive noise*

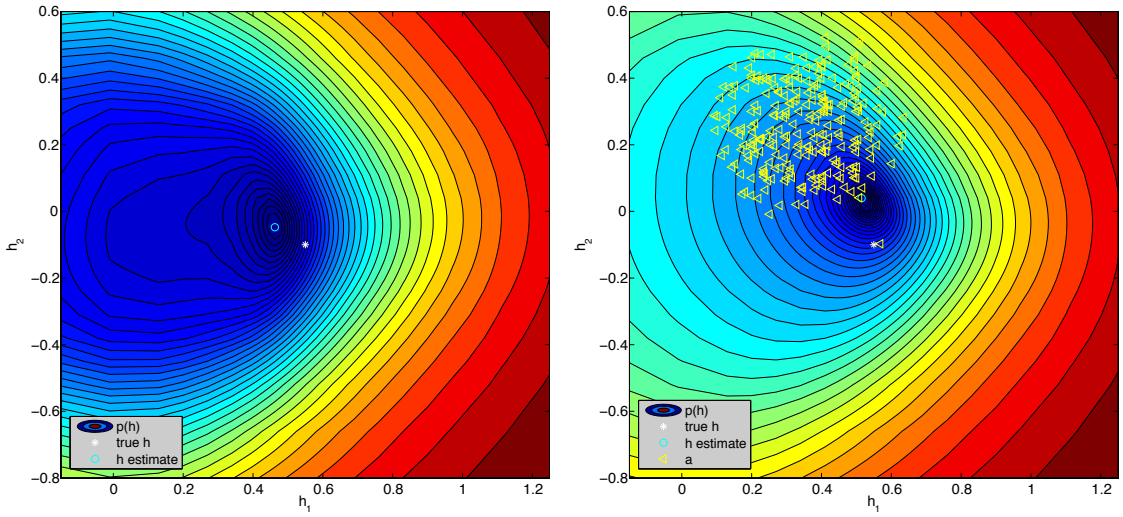
[6]Consider 30dB BSNR is a typical figure used in image restoration tests, however the AR blur being used here is of course unusual.



(a) as fig. 7.9(b) but with 5dB BSNR

(b) as fig. 7.8(b) but with 5dB BSNR

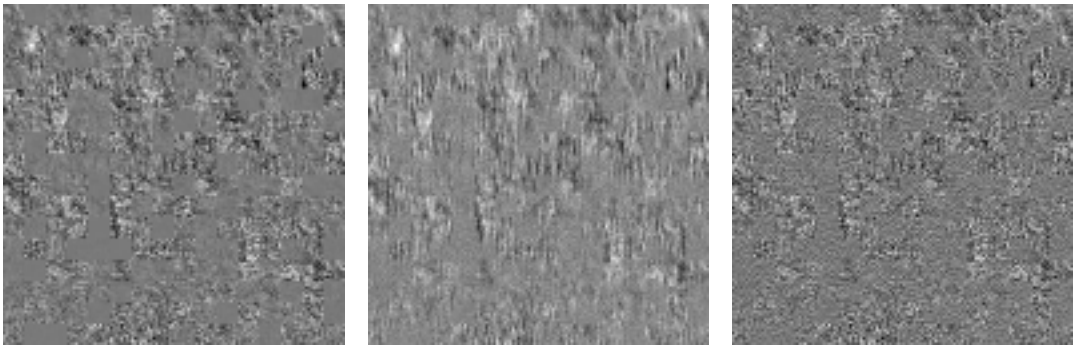
Figure 7.14: *Experiment 5. Effect of additive noise*



(a) as fig. 7.12(b) but with 5dB BSNR

(b) as fig. 7.11(a) but with 5dB BSNR

Figure 7.15: *Experiment 5. Effect of additive noise*



(a) $\mathbf{f}$ (b) $\mathbf{g}$, including additive WGN (c) $\hat{\mathbf{f}}$, restored using estimated blur

Figure 7.16: *Experiment 5. Images used in experiment with 10dB BSNR, 256 Block AR(1) source, estimated as AR(3)*

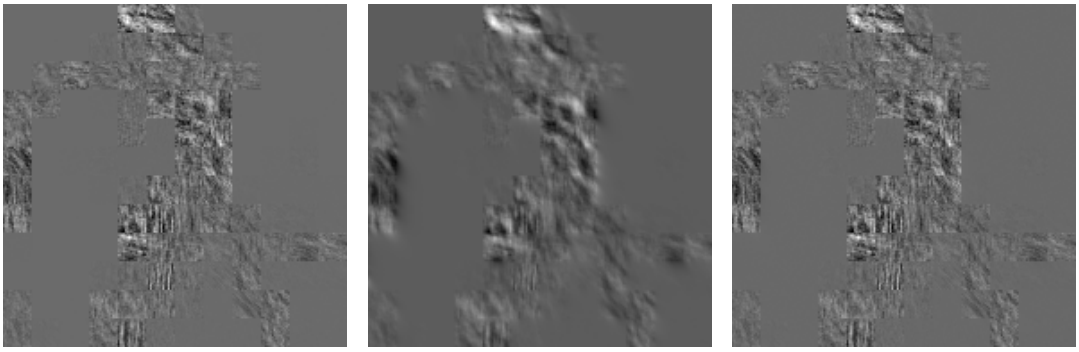
7.2.6 Experiment 6. — Synthetic image, AR blur

Having examined the effects on systematically generated data, we try an experiment with source parameters from a real image using an ML estimate. In the example shown in Fig. 7.17(a), a first order AR process is used for the blur and the source image. WGN is used to drive the source, whose AR parameters $\hat{\mathbf{a}}_r$ for each 16×16 pixel block are estimated from the 256×256 *Cameraman* image (in fig. 7.22(a)). Noise is added to the blurred image at 40dB BSNR (see Fig. 7.17(b)). For this experiment, we use $\mathbf{h}' = [0.45, 0.35, 0.05]$. In Fig. 7.18, the estimate of the blur parameters is seen to be correctly estimated as $[\hat{h}_1, \hat{h}_2] = [0.45, 0.35]$.

The experiment is repeated, but with optimisation across all three parameters of $\mathbf{h}'$, using the deterministic nonlinear Nelder-Mead Simplex method [118] (`fminsearch` in MATLAB), producing $\hat{\mathbf{h}}' = [0.442, 0.347, 0.058]$. The restored image using the inverse filter $(\mathbf{I} - \mathbf{H}')$ is shown in Fig. 7.17(c).

7.2.7 Experiment 7. — Real image, AR blur

The method is now tested with a real image as the source. Note that in this case, there are no ground truth AR parameters to compare to. An interesting effect also occurs during blur estimation using the MMAP estimate. Using knowledge acquired in the previous experiments, the estimation is performed with 8×8 blocks and source model of order 3. A 128-pixel image is shown in fig. 7.19(a), which is blurred by the same $\mathbf{h}$ as before, giving fig. 7.19(b). When deblurred with the estimate of fig. 7.21(a), the image of fig. 7.19(c) is obtained.



(a) $\mathbf{f}$ (b) $\mathbf{g}$ (c) $\hat{\mathbf{f}}$

Figure 7.17: *Experiment 6. (a) 1st order BSAR synthetic image with 16×16 pixel blocks (b) Blurred image, at 40dB BSNR (c) Restored image. (central portion shown)*

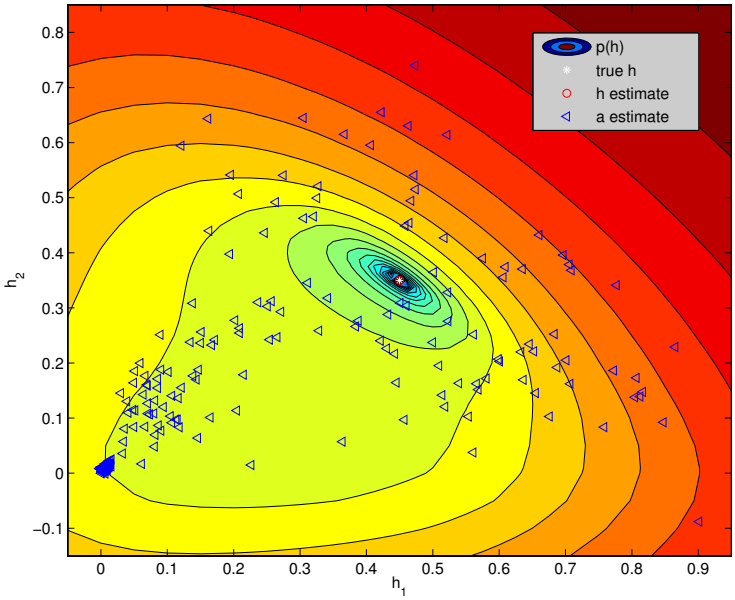


Figure 7.18: *Experiment 6.* $\mathrm{p}(\mathbf{h}'|\mathbf{g})$*, for the synthetic AR image.*

(a) $\mathbf{f}$ (b) $\mathbf{g}$, no added noise (c) $\hat{\mathbf{f}}$, restored using estimated blur

Figure 7.19: *Experiment 7. Real image blurred with AR blur, estimated as 256 Block AR(3) source*

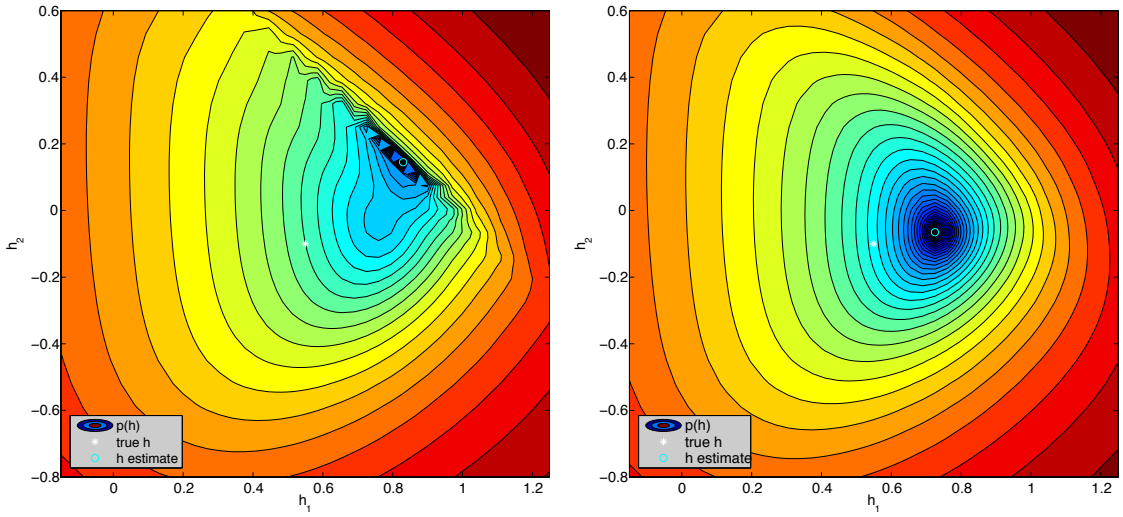
(a) $\mathbf{f}$, high pass filtered real image (b) $\mathbf{g}$, no added noise (c) $\hat{\mathbf{f}}$, restored using estimated blur

Figure 7.20: *Experiment 7. Real image, high pass filtered then blurred with AR blur, estimated as 256 Block AR(3) source*

The estimate is obviously a poor one, but there is a possible explanation for this. The low-order AR model does not very well capture both the local correlations and the sudden jumps at edges, or local-mean characteristics, of the true image; it is best suited to modelling zero-mean alternating signals such as audio signals, or texture in images.[7] Thus the model tries to find a filter that will 'restore' the image to this state as is successfully achieved in fig. 7.19(c): the high frequencies have been successfully restored and an image that best matches the AR model is obtained.

If instead we begin with a 'real' image that has had its local mean removed by high-pass

[7]Strictly speaking, the images are positive valued only, however the global mean was subtracted prior to processing. Even with this procedure, within each region the image possesses an inherently non-zero mean, or much of its energy concentrated in very low frequencies.



(a) Estimate of blur using real source image

(b) Estimate of blur using high-pass filtered real image for source

Figure 7.21: *Experiment 7. Probability for AR blur parameters for real images in Experiment 7*

filtering,[8] as in fig. 7.20(a), blur it similarly, fig. 7.20(b), and then the estimate of fig. 7.21 gives the deblurred image as in fig. 7.20(c). This estimate obviously resembles the source image more closely since it better fits the BSAR model in the first place. It can be seen that the estimate of the blur parameters is still not perfect, but better than with the previous image.

7.3 Incorporating local means

In order to make the estimation work with real images, some modification of the model is needed. Unlike true AR signals, real images are inherently non-zero mean. Removing the global mean from the image is not enough; the estimated local sample mean of each block is required to more accurately model the true image. Rather than re-derive the marginalised expression incorporating the means into the model, a simple change is proposed here, which was found to give correct experimental results. In the estimation procedure, after $\hat{\mathbf{f}}$ is found, the mean of each block is subtracted. Care is required when subtracting these means from the matrix $\hat{\mathbf{F}}$, since it is the local mean for each block that should be used; as such any data elements providing boundary conditions for a block must use the mean of their neighbour and

[8] Performed in this case simply by convolving the image with a 10-pixel circular blur to low-pass filter it, then subtracting this component from the original image.

not their own block mean, as shown in §6.4, and these will have different values depending on the block in question. This procedure avoids discontinuities in the data. An experiment using the empirically derived local mean method is now presented.

7.3.1 Experiment 8 — Real Image, Synthetic Blur, including means

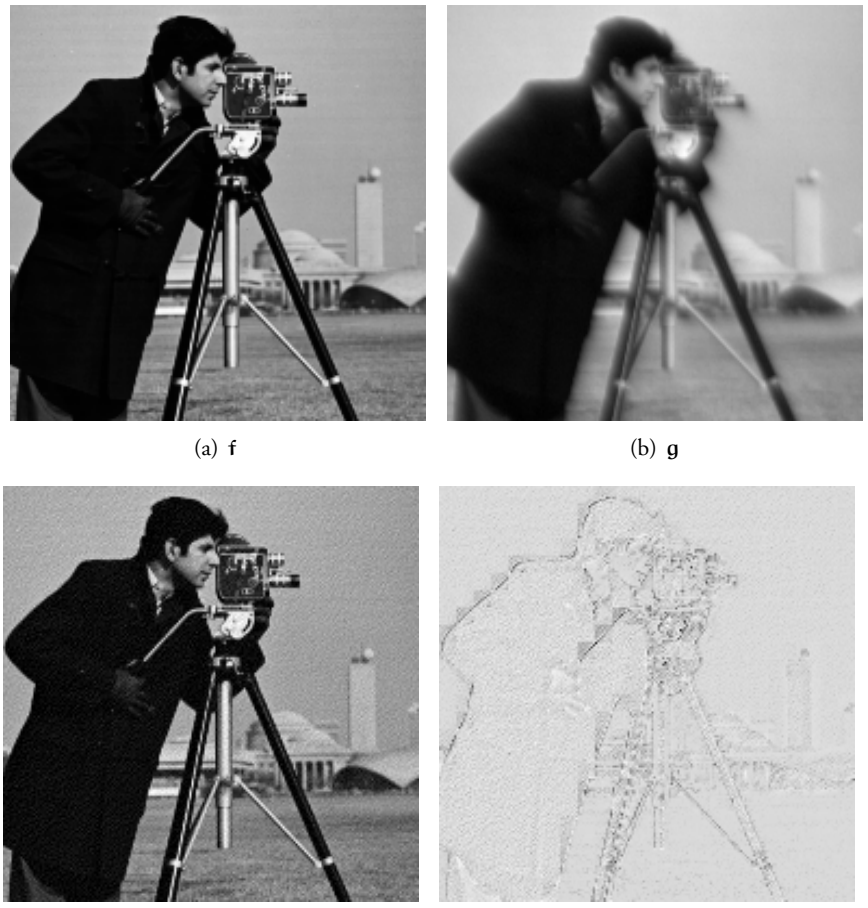
The 256 × 256 pixel *Cameraman* image is blurred with the same parameters as in experiment 6, and noise added at 35dB BSNR, shown in Fig. 7.22. The blur is estimated using the same deterministic optimisation method, using 8 × 8 pixel blocks, with hyperparameters $\alpha_r, \beta_r = 0, \delta_{\alpha_f} = 1$.

It has been seen that using higher than 1st order models can give better results, and here a 3rd order source model is used. With these changes — despite exclusion of the noise from the model — a successful restoration is still possible. At lower BSNRs, the restoration is still sharp, but the amplified noise becomes more prominent.
The estimated blur is found as $\hat{\mathbf{h}}' = [0.473, 0.377, 0.011]$ in the 35dB case. Due to the slight error, a small number of super-white and super-black pixels are produced with intensities outside the range of the source image (giving longer tails in the image histogram). Thus to display the image correctly, these tails should be clipped, or histogram specification may be used. The image estimate after clipping is shown in Fig. 7.22(c). The model fitting term (§7.1.6) is shown in Fig. 7.22(d).

7.4 Discussion: Using Non-causal Models

Some attempts have been made to extend the MMAP estimator to work with more realistic non-causal blur models — the model could also be extended to non-causal image models, but the blur is the main factor holding the method back from being applied to real degradations. The main issue is the determinant $\det|\mathbf{H}|$ that appears when the model is non-causal.

Evaluation of $\det|\mathbf{H}|$ may cause problems due to the large size of the matrix (numerical underflow may occur if the log is not taken) — the same splitting procedure cannot be used as it is not block diagonal, rather it is BTTB. A matrix factorization (either $\mathbf{LU}$ or $\mathbf{QR}$, or Cholesky if the PSF is symmetric) may be used, with the log-determinant calculated from the sum of the diagonal terms of the factors (the $\mathbf{U}$ or $\mathbf{R}$ matrices are upper triangular, while the $\mathbf{L}$ or $\mathbf{Q}$



(a) $\mathbf{f}$ (b) $\mathbf{g}$

(c) $\hat{\mathbf{f}}$ (d) $\hat{\mathbf{f}}^{\mathsf{T}}(\hat{\mathbf{f}} - \hat{\mathbf{A}}\hat{\mathbf{f}})$

Figure 7.22: *Experiment 8. Cameraman (a) source; (b) blurred; (c) restored and (d) weighted prediction error images. Central portion shown.*

matrices have unity determinant and can be ignored). In practice to enable efficient calculation, it may be necessary to use the BCCB approximation of §3.2.1 (i.e. that the convolution is circular), which enables the sum of the coefficients of the 2D-FFT of the PSF to be used instead.

Due to the approximations made, namely ignoring the noise variance, many singularities appear in the blur parameter space where the determinant is zero. If some regularisation were used, these zeros could be smoothed out. Unfortunately, doing this would make marginalisation of the other parameters analytically intractible. However, there is a more fundamental problem to be considered. If there is no noise, then we should be able to obtain a reliable estimate. This is not the case, due to the original marginalisation of the image.

Because the blur itself may be singular, the original probability transformation from $\mathbf{f}$ to $\mathbf{g}$ is underdetermined — there is a degeneracy in the likelihood and the space of possible blurred images lies on a manifold. When there are a multiplicity of solutions, the Jacobian cannot be expressed simply as the determinant of the singular matrix $\mathbf{H}$, but we should consider summing over the Jacobians evaluated at all the possible images in the subspace that map on to this observation. This means the marginalised form of the likelihood becomes a non-trivial way to treat the problem in these cases, and we are better to work with the complete likelihood.

These problems are partly related to the modelling of the boundary conditions, which is the subject of the next chapter.

7.5 Conclusions

This chapter has investigated estimation of PSF parameters using the MMAP method. This method allows direct estimation of these parameters via maximisation of the marginalised posterior, i.e. the distribution of $\mathrm{p}(\mathbf{h}'|\mathbf{g})$. The marginalisation avoids having to estimate any of the other model parameters, but still takes into account uncertainty in their values.

This has provided an interesting way to study the BID problem, and in particular how different settings of the proposed image model affect the reliability of the results. Unfortunately, due to approximations that have been made in order to derive the marginalised posterior, this method cannot be used for the estimation of general realistic PSFs, which are noncausal, that we need to properly treat blurs such as defocus. In Chapters 9 and 10, we consider other Bayesian methods using the complete (non-marginalised) model which do not suffer from

these problems, after first taking a closer look at boundary modelling.

Chapter 8
Boundary Modelling

In most image processing problems, it is assumed that the observation region is the same size as the region of the underlying data, i.e. data from outside the boundary does not affect the observation.

In practice, because the image is only recorded within a finite region on the image plane, and the real world has an effectively infinite extent, information from outside the observed region will blur inside and contaminate the observation. Thus any attempt to restore the image without taking account of these effects will produce erroneous results. Typically this will manifest as large ringing errors propagating from the image boundaries.

For some applications — medical imaging and astronomy in particular — it is true that the object that is being imaged is contained inside a finite support, outside of which there is zero intensity. In these cases the zero boundary conditions just described are appropriate, and methods that rely on this — for instance, naïve Fourier processing which implies a circulant image boundary, or methods that enforce the zero boundary such as nonnegativity and support constraints with recursive image filtering (NAS-RIF) [117] — can be used without issues.

8.1 Methods for suppressing boundary artefacts

Typically when dealing with natural photographic images however, ad-hoc artefact reduction methods have been employed. Because the desire to use Fourier domain restoration exploiting the fast Fourier transform (FFT) has been very strong when faced with limited computational resources, methods are used to *pre-process* the borders to make the blurred image approximately follow the circulant model. One method is to create artificial reflective, or *Neumann* boundary conditions (making the derivative zero across the boundary), by enlarging the image with a mirrored version [3]. Other methods have been proposed that actually assume a reflected boundary model outside the image, and use preconditioning techniques to restore according to this model [162], or "anti-reflective" boundaries for higher order smoothness [197].

Alternatively, smoothing using a window function or linear interpolation, and zero-padding can be employed to blend values from one side of the image into the other [120], at the expense of reduction of the restored image size.

A final technique is when using an iterative spatial-domain restoration procedure, to weight the boundaries such that they are treated as progressively more unreliable data close to the edge, and hence the edge regions do not converge as quickly and do not contribute to the restoration. For example, the iteration of (3.9) is used with $\mathbf{R}$ tapering from towards zero near the unreliable edge pixels, so that they converge more slowly. Using nonlinear POCS with this method also helps stop the oscillations from the modelling errors from building up and swamping the restoration. Although reasonably successful in suppressing artefacts, this technique requires choosing extra parameters that don't correspond to the model.

One reason why the artefacts tend to be difficult to remove in this manner is because although the support of the blur is small, the support of the inverse filter (or the restoration filter) tends to be very large. Consequentially, any errors in the model at the boundary tend to propagate across the image, even if the blur is known exactly.

All these methods tend to just suppress the artefacts due to an incorrect boundary model being used, rather than avoid them, and do not make full use of the available data. In this chapter we consider the correct way to treat the boundaries under the Bayesian framework, that is to treat the image data outside the observation region as a missing data problem.

8.1.1 Examples

In fig. 8.2, some of the methods just described for artefact suppression are shown, and compared with the extended estimation model that will be explained in this chapter. In fig. 8.1, the central part of the Lena image is shown; this is blurred firstly with a small Gaussian PSF of standard deviation $\sigma = 2$ pixels, and noise added at 60dB BSNR (equal to a variance of 3×10^{-3} gray levels), and secondly a Gaussian of standard deviation $\sigma = 3$ and noise at 40dB (a variance of 0.3 levels). The first case is used for the restorations in fig. 8.2(a) and 8.2(b) and the second in fig. 8.2(c) and 8.2(d).

The restorations within each figure show the effects of the different methods, all based on a restoration using the simple stationary SAR prior with the discrete Laplacian, i.e. estimating $\hat{\mathbf{f}}$ from (3.4). In each sub-figure, the boundary conditions in the six restorations from left

to right, then top to bottom, are: (i) Fourier domain solution, no pre-processing (assumed periodic boundary conditions); (ii) Fourier solution, pre-processed with linear interpolation of borders [1]; (iii) Fourier solution, pre-processed to conform with the reflected (Neumann) boundary condition[2]; (iv) CG solution, 500 iterations (assumed zero, or Dirichlet, boundary conditions); (v) Iterative Tikhonov solution using POCS and boundary weighting, 3000 iterations; (vi) CG solution using extrapolation and extended boundary observation model (to be described in §8.3), 500 iterations.

Each set of six images has a different amount of regularisation (more can be seen to be required for the larger blur/lower BSNR), which reveals the effects this has on the artefacts. Whilst with a larger amount of regularisation, the edge interpolation and reflection methods can suppress most of the artefacts, only the proposed method succeeds in eliminating them, and the iteratively weighted POCS method comes a close second with some very minor loss of detail at the edge. The intensities of the restorations are clipped to the range 0–255 to better see the detail in the images, as the oscillations near the edges grow rapidly in intensity.

Figure 8.1: *Images used for boundary artefact examples. From left to right: original image; 2 pixel Gaussian blur at 60dB; 3 pixel Gaussian blur at 40dB*

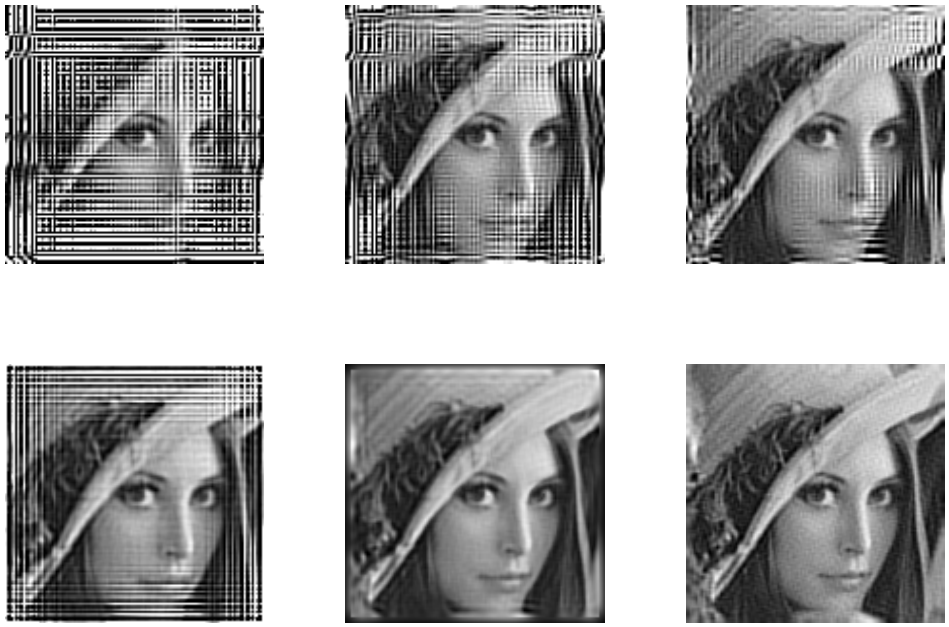
8.1.2 Related work

While the methods presented in this chapter were under investigation, it was discovered that similar research had just been published by Calvetti & Somersalo [37–39]. The end results are the same, although the derivation presented here explicitly uses the Bayesian marginalisation approach, whereas Calvetti & Somersalo apply results for the conditional mean of partitioned

[1] Image is extrapolated using reflected boundary, then a small portion is copied from one side to the other with linear blending of the two towards the edge. The image size is maintained.

[2] Image is tiled to 4 times original size, such that the central part is the original image and the image outside the original boundary is its reflection. This larger tile is then periodic.

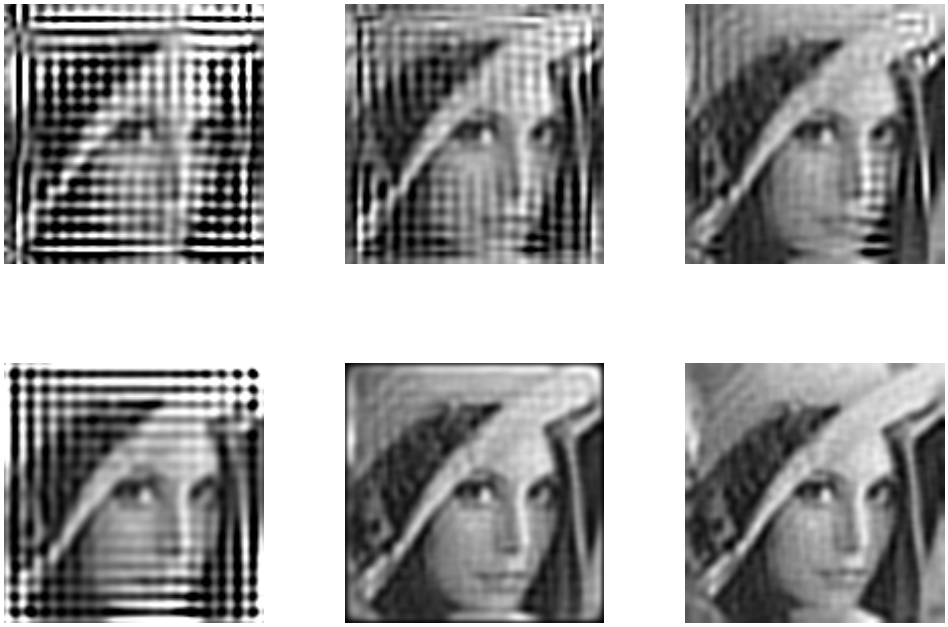
(a) $\sigma = 2$ pixel Gaussian blur (12 pixel support), 60dB BSNR, Regularisation parameter $\alpha = 10^{-2}$



(b) $\sigma = 2$ pixel Gaussian blur (12 pixel support), 60dB BSNR, Regularisation parameter $\alpha = 10^{-5}$

Figure 8.2: *Restorations using different boundary conditions, all with same SAR image prior. See main text in §8.1.1 for a description of the different restorations in each case.*

(c) $\sigma = 3$ pixel Gaussian blur (12 pixel support), 40dB BSNR, Regularisation parameter $\alpha = 10^{-1}$



(d) $\sigma = 3$ pixel Gaussian blur (12 pixel support), 40dB BSNR, Regularisation parameter $\alpha = 10^{-2}$

Figure 8.2: *cont.*

Gaussian distributions directly. The two results may be related after some extensive algebraic manipulation, as shown in Appendix E.

8.2 Extended Observation Model

The key to correctly avoiding boundary artefacts is to re-write the observation model to take account of the original image outside the observed region, i.e. to treat it as a *missing data* problem.

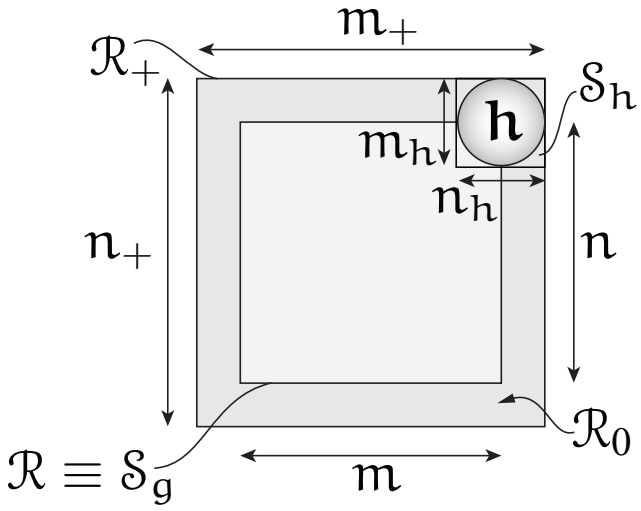


Figure 8.3: *The extended image region $\mathcal{R}_+$, boundary region $\mathcal{R}_0$, and observation region $\mathcal{R}$. $\mathcal{R}_+$ is of size $\{m_+, n_+\} = \{m + m_h - 1, n + n_h - 1\}$, since the PSF is assumed to be zero outside a support $\mathcal{S}_h$ of size $m_h \times n_h$.*

To facilitate this, we define the extended observation model using the regions shown in Fig. 8.3. Assume that the true image is defined on an infinite canvas Ω_s (as described in §2.1), which we will partition into some specific regions. Firstly, we define the *observation region* $\mathcal{R}$ equal to $\mathcal{S}_g$, corresponding to the support of the observed image $\mathbf{g}$; this is fixed by the physical camera model. $\mathcal{R}$ contains $L = mn$ pixels. Then we define a *boundary region* $\mathcal{R}_0$ around this. Pixels of the true image in $\mathcal{R}_0$ influence the blurred image in $\mathcal{R}$ — that is, the size of the boundary is determined by the blur support $\mathcal{S}_h$. The *extended region* $\mathcal{R}_+$ is the union of $\mathcal{R}$ and $\mathcal{R}_0$, and contains $L_+ = m_+ n_+$ pixels.

We are free to choose the support $\mathcal{S}_f$ of the true image $\mathbf{f}$ in the observation model to be any size we want; commonly it is taken to be $\mathcal{R}$, i.e. the same as $\mathcal{S}_g$. However it should be clear from the discussion that letting $\mathcal{S}_f$ equal $\mathcal{R}_+$ is the only choice that takes account properly of

all the pixels blurring into the observed frame. To make this distinction clear, for the duration of this chapter, we will rewrite the true image with support $\mathcal{S}_\mathbf{f}$ on $\mathcal{R}_+$ as $\mathbf{f}_+$. We then partition the vector $\mathbf{f}_+$ into $\mathbf{f}_1$ defined on $\mathcal{R}$, and $\mathbf{f}_0$, defined on the boundary region $\mathcal{R}_0$. Note that $\mathbf{f}_1$ in this case is just the same as $\mathbf{f}$ in the regular case where $\mathcal{S}_\mathbf{f}$ is the same as $\mathcal{S}_\mathbf{g}$, though the distinction is made because $\mathbf{f}_1$ is only defined on a subset of $\mathcal{S}_\mathbf{f}$ in the extended case. With this notation, the extended observation model is written

$$\mathbf{g} = \mathbf{H}_+\mathbf{f}_+ + \boldsymbol{w} = \left[\begin{array}{c|c} \mathbf{H}_1 & \mathbf{H}_0 \end{array}\right] \left[\frac{\mathbf{f}_1}{\mathbf{f}_0}\right] + \boldsymbol{w} \tag{8.1}$$

$$= \mathbf{H}_1\mathbf{f}_1 + \mathbf{H}_0\mathbf{f}_0 + \boldsymbol{w}. \tag{8.2}$$

Given the extended observation model, there are two ways we may obtain an estimate of $\mathbf{f}_1$. Firstly, we may estimate the extended image $\mathbf{f}_+$, which corresponds to extrapolation based on the available data. We may then simply discard the estimated boundary pixels (or keep them if they prove useful). Secondly, we may marginalise the boundary pixels $\mathbf{f}_0$, as they are not strictly required for the restoration, and just estimate $\mathbf{f}_1$ directly. These options will be considered in the following two sections.

In either case, the important factor is that to properly complete the Bayesian model, we must place priors over all unknown variables that influence the observation. Hence the prior for the true image must be specified for $\mathbf{f}_+$.

It should be clear that (8.1) is always an underdetermined system as $\mathbf{H}_+$ is wider than it is tall. Thus without a good prior we can never hope to recover the true image, even in the noiseless case. If the nullspace of the blur does not significantly intersect the space of true images we expect to see, then with a good prior that distinguishes the two we should be able to recover the image well in the noiseless case. This is similar to the problem when the blur itself has zeros in its spectra. In fact the boundary conditions have an important influence over the recoverability. Typical analysis considers the DFT of the PSF, and regards frequencies that are zero as being unrecoverable. This is an approximation, which only holds under circulant boundaries. To see the full picture, the singular value decomposition (SVD) of the extended blur matrix $\mathbf{H}_+$ should be examined. It is more likely that the matrix is ill-conditioned rather than actually singular, in which case the important factor is the rate of decay of its singular values — having more small singular values will make the image less recoverable in general.

8.3 Extrapolation

We will first consider the case where we estimate the extended image $\mathbf{f}_+$ in full. Assume that we wish to restore the image based upon the observations $\mathbf{g}$ under the model in (8.1), but all other parameters controlling the restoration are known (we denote these by a $\cdot$). To estimate the posterior distribution of the image $\mathbf{f}_+$ in the Bayesian framework, we need a prior $p(\mathbf{f}_+ \mid \cdot)$, which in general we can take to be $\mathcal{N}\left(\mathbf{f}_+ \,\middle|\, \bar{\mathbf{f}}_+, \boldsymbol{\Sigma}_{\mathbf{f}_+}\right)$. We will assume here $\bar{\mathbf{f}}_+ = 0$ for simplicity, as this is the common case [3]. The posterior for the extended image is then found as

$$p(\mathbf{f}_+ \mid \mathbf{g}, \cdot) \propto p(\mathbf{g} \mid \mathbf{f}_+, \cdot)\, p(\mathbf{f}_+ \mid \cdot) \tag{8.3}$$

$$= \exp\Big(-\frac{1}{2}\big[\mathbf{f}_+^{\mathsf{T}}(\mathbf{H}_+^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{H}_+ + \boldsymbol{\Sigma}_{\mathbf{f}_+}^{-1})\mathbf{f}_+ - 2\mathbf{f}_+^{\mathsf{T}}(\mathbf{H}_+^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{g}) + \mathbf{g}^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{g}\big]\Big) \tag{8.4}$$

which is a Gaussian

$$p(\mathbf{f}_+ \mid \mathbf{g}) \propto \mathcal{N}\left(\mathbf{f}_+ \,\middle|\, \boldsymbol{\mu}_{\hat{\mathbf{f}}_+}, \boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_+}\right) \tag{8.5}$$

$$= \exp\Big(-\frac{1}{2}\big[\mathbf{f}_+^{\mathsf{T}}\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_+}^{-1}\mathbf{f}_+ - 2\mathbf{f}_+^{\mathsf{T}}\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_+}^{-1}\boldsymbol{\mu}_{\hat{\mathbf{f}}_+} + \boldsymbol{\mu}_{\hat{\mathbf{f}}_+}^{\mathsf{T}}\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_+}^{-1}\boldsymbol{\mu}_{\hat{\mathbf{f}}_+}\big]\Big). \tag{8.6}$$

By comparison of (8.4) and (8.6), the parameters are given as

$$\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_+}^{-1} = \mathbf{H}_+^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{H}_+ + \boldsymbol{\Sigma}_{\mathbf{f}_+}^{-1} \tag{8.7}$$

$$\boldsymbol{\mu}_{\hat{\mathbf{f}}_+} = \boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_+}\left(\mathbf{H}_+^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{g}\right). \tag{8.8}$$

We can take the MAP estimate as $\hat{\mathbf{f}}_+ = \boldsymbol{\mu}_{\hat{\mathbf{f}}_+}$. Note the similarity to (4.19) for the standard observation model where $\mathcal{S}_\mathbf{f} = \mathcal{R}$: we have just used the non-square matrix $\mathbf{H}_+$ instead of $\mathbf{H}$, along with the extended prior. In the examples section it can be seen that this estimate corresponds to extrapolating the image at the boundaries. Some of the extra recovered information comes from the blurred image, and some from the prior.

8.4 Boundary Marginalisation

The alternative case to consider is when we only estimate the image in $\mathcal{R}$ and may not care about the actual value of the boundary pixels $\mathbf{f}_0$. However we need to take into account the

[3] If we are estimating a tile from the image and already have an estimate of another part, we may have some extra information about the expected value of the boundaries; this may be incorporated through the relevant subset of pixels in $\bar{\mathbf{f}}_+$

uncertainty in their value, by placing the prior on the full extended region, as before, and then find the posterior for the desired region alone, $\mathbf{f}_1$, by integrating out the boundary region $\mathbf{f}_0$:

$$\mathrm{p}\left(\mathbf{f}_1 \,|\, \mathbf{g}, \cdot\right) = \int \mathrm{p}\left(\mathbf{f}_+ \,|\, \mathbf{g}, \cdot\right) \mathrm{d}\mathbf{f}_0 = \int \mathrm{p}\left(\mathbf{g} \,|\, \mathbf{f}_+, \cdot\right) \mathrm{p}\left(\mathbf{f}_+ \,|\, , \cdot\right) \mathrm{d}\mathbf{f}_0. \tag{8.9}$$

We can substitute the posterior from (8.4), but to perform the marginalisation, we will partition the prior $\mathrm{p}\left(\mathbf{f}_+ \,|\, , \cdot\right) = \mathcal{N}\left(\mathbf{f}_+ \,\middle|\, \bar{\mathbf{f}}_+, \boldsymbol{\Sigma}_{\mathbf{f}_+}\right)$ according to

$$\bar{\mathbf{f}}_+ = \left[\begin{array}{c} 0 \\ \hline \bar{\mathbf{f}}_0 \end{array}\right] \qquad \text{and} \qquad \boldsymbol{\Sigma}_{\mathbf{f}_+} = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{\mathbf{f}_1}^{-1} & \boldsymbol{\Sigma}_{\mathbf{f}_1,\mathbf{f}_0}^{-1} \\ \hline \boldsymbol{\Sigma}_{\mathbf{f}_1,\mathbf{f}_0}^{-\mathsf{T}} & \boldsymbol{\Sigma}_{\mathbf{f}_0}^{-1} \end{array}\right]^{-1}. \tag{8.10}$$

where we now have allowed for information on the mean of the boundary to be specified. The marginalisation in (8.9) may then be found using the Gaussian integral; details of the derivation are given in Appendix E. The result is another Gaussian,

$$\mathrm{p}\left(\mathbf{f}_1 \,|\, \mathbf{g}\right) \propto \mathcal{N}\left(\mathbf{f}_1 \,\middle|\, \boldsymbol{\mu}_{\hat{\mathbf{f}}}, \boldsymbol{\Sigma}_{\hat{\mathbf{f}}}\right) \tag{8.11}$$

with parameters

$$\boldsymbol{\Sigma}_{\hat{\mathbf{f}}}^{-1} = \mathbf{H}_1^{\mathsf{T}} \mathbf{Q}_w^{-1} \mathbf{H}_1 + \boldsymbol{\Sigma}_{\mathbf{f}_1}^{-1} - \boldsymbol{\Sigma}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}}^{-1} \boldsymbol{\Sigma}_{\hat{\mathbf{f}}_0} \boldsymbol{\Sigma}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}}^{-\mathsf{T}} \tag{8.12}$$

$$\boldsymbol{\mu}_{\hat{\mathbf{f}}} = \boldsymbol{\Sigma}_{\hat{\mathbf{f}}} \Big(\Big(\mathbf{H}_1^{\mathsf{T}} \mathbf{Q}_w^{-1} - \boldsymbol{\Sigma}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}}^{-1} \boldsymbol{\Sigma}_{\hat{\mathbf{f}}_0} \mathbf{H}_0^{\mathsf{T}} \mathbf{Q}_w^{-1} \Big) \mathbf{g} + \Big(\boldsymbol{\Sigma}_{\mathbf{f}_1,\mathbf{f}_0}^{-1} - \boldsymbol{\Sigma}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}}^{-1} \boldsymbol{\Sigma}_{\hat{\mathbf{f}}_0} \boldsymbol{\Sigma}_{\mathbf{f}_0}^{-\mathsf{T}} \Big) \bar{\mathbf{f}}_0 \Big). \tag{8.13}$$

where we have defined $\boldsymbol{\Sigma}_{\hat{\mathbf{f}}_0} = (\mathbf{H}_0^{\mathsf{T}} \mathbf{Q}_w^{-1} \mathbf{H}_0 + \boldsymbol{\Sigma}_{\mathbf{f}_0}^{-1})^{-1}$ and $\boldsymbol{\Sigma}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}}^{-1} = (\mathbf{H}_1^{\mathsf{T}} \mathbf{Q}_w^{-1} \mathbf{H}_0 + \boldsymbol{\Sigma}_{\mathbf{f}_1,\mathbf{f}_0}^{-1})$.

We can now examine how this solution compares to the others. When the cross-terms of the image model prior — $\boldsymbol{\Sigma}_{\mathbf{f}_1,\mathbf{f}_0}^{-1}$, representing the correlation between the pixels at the edge of the desired region $\mathbf{f}_1$ and the boundary region $\mathbf{f}_0$ — and the boundary prior covariance itself, $\boldsymbol{\Sigma}_{\mathbf{f}_0}$, are zero, the solution reduces to the standard restoration for the desired region. Specifically, $\boldsymbol{\Sigma}_{\hat{\mathbf{f}}}$ becomes just $\mathbf{H}_1^{\mathsf{T}} \mathbf{Q}_w^{-1} \mathbf{H}_1 + \boldsymbol{\Sigma}_{\mathbf{f}_1}^{-1}$ and hence $\boldsymbol{\mu}_{\hat{\mathbf{f}}} = \left(\mathbf{H}_1^{\mathsf{T}} \mathbf{Q}_w^{-1} \mathbf{H}_1 + \boldsymbol{\Sigma}_{\mathbf{f}_1}^{-1} \right)^{-1} (\mathbf{H}_1^{\mathsf{T}} \mathbf{Q}_w^{-1} \mathbf{g})$. These assumptions are implying that firstly the image boundary is uncorrelated with the image inside the observed window, and that the boundary conditions are known exactly, which is rarely the case in reality.

Typically it will be natural to assume that the mean of the boundary, $\bar{\mathbf{f}}_0$, is unknown and hence will be set to zero. However it may be used if we are restoring an image block by block,

whereby the preceding estimate from one block could be used as the prior boundary mean for a neighbouring block to be deblurred, with a small boundary covariance allowing for the uncertainty/error in the previous estimate.

We can also examine a generalised case of the marginalisation, where the pixels that are marginalised are not exactly those in $\mathcal{R}_0$. Thus we let $\mathbf{f}_0$ be the vector of pixels we wish to keep, and $\mathbf{f}_1$ those we will estimate. Then the calculations proceed as before, and the only modification is to change the partitioning of the prior and $\mathbf{H}_+$ accordingly. Clearly if we let the partitioning tend to zero pixels in $\mathbf{f}_0$, the extraneous terms due to marginalisation in (8.13) disappear and we recover (8.8) as in the extended case. In §8.7, we will also examine intermediate cases where we marginalise only some of the boundary pixels, as something of a compromise between the two solutions.

8.5 Deterministic boundary prior

We can examine what happens when the boundary is known exactly. This implies independence of $\mathbf{f}_1$ and $\mathbf{f}_0$, i.e.

$$\mathrm{p}\left(\mathbf{f}_+ \,|\, \mathbf{g}\right) \propto \mathrm{p}\left(\mathbf{g} \,|\, \mathbf{f}_0, \mathbf{f}_1\right) \mathrm{p}\left(\mathbf{f}_1 \,|\, \boldsymbol{\Sigma}_{\mathbf{f}_1}\right) \mathrm{p}\left(\mathbf{f}_0 \,\middle|\, \bar{\mathbf{f}}_0\right) \tag{8.14}$$

and the Gaussian prior for $\mathbf{f}_0$ in the previous section can be replaced by a delta function, i.e.

$$\mathrm{p}\left(\mathbf{f}_0 \,\middle|\, \bar{\mathbf{f}}_0\right) = \delta(\mathbf{f}_0 - \bar{\mathbf{f}}_0). \tag{8.15}$$

Now when the result of integrating over $\mathbf{f}_0$ (substituting (E.6) and (8.15) into (8.14)) is:

$$\mathrm{p}\left(\mathbf{f}_1 \,|\, \mathbf{g}\right) \propto \int \mathrm{p}\left(\mathbf{g} \,|\, \mathbf{f}_+\right) \mathrm{p}\left(\mathbf{f}_+ \,\middle|\, \boldsymbol{\Sigma}_{\mathbf{f}_+}\right) \,\mathrm{d}\mathbf{f}_0 \tag{8.16}$$

$$\propto \mathrm{p}\left(\mathbf{f}_1 \,|\, \boldsymbol{\Sigma}_{\mathbf{f}_1}\right) \int \mathrm{p}\left(\mathbf{g} \,|\, \mathbf{f}_0, \mathbf{f}_1\right) \mathrm{p}\left(\mathbf{f}_0 \,\middle|\, \bar{\mathbf{f}}_0\right) \,\mathrm{d}\mathbf{f}_0 \tag{8.17}$$

$$\propto \mathrm{p}\left(\mathbf{f}_1 \,|\, \boldsymbol{\Sigma}_{\mathbf{f}_1}\right) \exp\Bigg[-\frac{1}{2}\Big(\mathbf{f}_1^{\mathsf{T}}(\mathbf{H}_1^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{H}_1)\mathbf{f}_1 - 2\mathbf{f}_1^{\mathsf{T}}\left(\mathbf{H}_1^{\mathsf{T}}\mathbf{Q}_w^{-1}(\mathbf{g} - \mathbf{H}_0\bar{\mathbf{f}}_0)\right) + \left(\mathbf{g}^{\mathsf{T}} - \mathbf{H}_0\bar{\mathbf{f}}_0\right)^{\mathsf{T}}\mathbf{Q}_w^{-1}\left(\mathbf{g}^{\mathsf{T}} - \mathbf{H}_0\bar{\mathbf{f}}_0\right)\Big)\Bigg], \tag{8.18}$$

by the sifting property of the delta function. Given the same initial prior for $\mathbf{f}_0$, this is of the

same form as (4.17) with

$$\boldsymbol{\Sigma}_{\hat{\mathrm{f}}}^{-1} = \mathbf{H}_1^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{H}_1 + \boldsymbol{\Sigma}_{\mathrm{f}_1}^{-1} \tag{8.19}$$

$$\boldsymbol{\mu}_{\hat{\mathrm{f}}} = \boldsymbol{\Sigma}_{\hat{\mathrm{f}}}\left(\mathbf{H}_1^{\mathsf{T}}\mathbf{Q}_w^{-1}(\mathbf{g}^{\mathsf{T}} - \mathbf{H}_0\bar{\mathbf{f}}_0)\right). \tag{8.20}$$

This is to be expected because if we set $\bar{\mathbf{g}} = \left(\mathbf{g}^{\mathsf{T}} - \mathbf{H}_0\bar{\mathbf{f}}_0\right)$ then we have just recovered the original problem without the extended boundary conditions, i.e. where the data $\bar{\mathbf{g}}$ really does come from a model with zero boundary conditions, and using the standard restoration model is actually appropriate.

However this solution should be equal to that found in the previous section assuming that $\boldsymbol{\Sigma}_{\mathrm{f}_0}$ and $\boldsymbol{\Sigma}_{\mathrm{f}_1,\mathrm{f}_0}^{-1}$ tend to zero (implying the Gaussian tends to a delta function at $\bar{\mathbf{f}}_0$).

8.6 Implementation details

In practice we generally wish to avoid explicit matrix inversion to find the solution to the mean of the posterior. Because the matrices involved are generally sparse, for images under about 100 pixels in each dimension, it becomes feasible with a moderate amount of memory (up to several hundred megabytes depending on the extent of the blur) to form the covariance matrix and apply standard direct linear algebra solvers such as Gaussian elimination (using MATLAB notation, $\boldsymbol{A}^{-1}\boldsymbol{x}$ becomes $\boldsymbol{A}\backslash\boldsymbol{x}$). For larger sized problems, it becomes more computationally efficient to apply iterative solvers such as CGLS.

8.6.1 Extended image estimation

For the extended image case, CGLS can be very efficient,. Because of the form of the posterior, we can repose the problem of estimating $\boldsymbol{\mu}_{\hat{\mathrm{f}}_+}$ in (8.8) as solving a linear system

$$\mathbf{M}^{\mathsf{T}}\mathbf{M}\boldsymbol{\mu}_{\hat{\mathrm{f}}_+} = \mathbf{M}^{\mathsf{T}}\mathbf{y} \tag{8.21}$$

where we let $\mathbf{M}^{\mathsf{T}}\mathbf{M} = \boldsymbol{\Sigma}_{\hat{\mathrm{f}}_+}^{-1}$ and $\boldsymbol{\Sigma}_{\mathrm{f}_+}^{-1} = \mathbf{C}^{\mathsf{T}}\mathbf{C}$, and assume $\mathbf{Q}_w = \sigma_w^2\mathbf{I}$; therefore

$$\mathbf{M} = \begin{bmatrix} \sigma_w^{-1}\mathbf{H}_+ \\ \hline \mathbf{C} \end{bmatrix}, \qquad \text{and} \qquad \mathbf{y} = \begin{bmatrix} \sigma_w^{-1}\mathbf{g} \\ \hline \mathbf{0}_{\mathrm{L}_+} \end{bmatrix}. \tag{8.22}$$

At each iteration main computational burden involves applying the matrix $\mathbf{M}$ and its transpose to a vector, without needing to form $\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_+}^{-1}$ in full. $\mathbf{H}_+$ is applied efficiently using an FFT based convolution (padded to avoid wraparound), whilst $\mathbf{C}$ is usually a very sparse banded matrix that is specified by the image model and can be applied directly with minimum effort.

8.6.2 Marginalised image estimation

The marginalised computations for evaluating (8.13) require some extra work, because of the additional term $-\boldsymbol{\Sigma}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}}^{-1}\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_0}\boldsymbol{\Sigma}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}}^{-\mathsf{T}}$ appearing in the covariance. The problem is to find $\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_0}$, which is the inverse of $\mathbf{H}_0^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{H}_0 + \boldsymbol{\Sigma}_{\mathbf{f}_0}^{-1}$. Fortunately the dimension is only equal to the number of pixels in the marginalised boundary $\mathcal{R}_0$ — typically much less than those in the observation region, so direct inversion is possible. However for efficiency and numerical stability, it is again better to avoid this. By instead performing a Cholesky decomposition of $\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_0}^{-1}$ as $\mathbf{K}^{\mathsf{T}}\mathbf{K}$, several other options become available.

Finding $\mathbf{K}^{-1}$ is quite efficient due to the triangular form of $\mathbf{K}$ and the sparsity of the original matrix. We can then find $\boldsymbol{\Sigma}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}}^{-1}\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_0}\boldsymbol{\Sigma}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}}^{-\mathsf{T}}$ as either $(\mathbf{K}^{\mathsf{T}}\backslash\boldsymbol{\Sigma}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}}^{-\mathsf{T}})^{\mathsf{T}}(\mathbf{K}^{\mathsf{T}}\backslash\boldsymbol{\Sigma}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}}^{-\mathsf{T}})$, $\boldsymbol{\Sigma}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}}^{-1}(\mathbf{K}^{-1}\mathbf{K}^{-\mathsf{T}})\boldsymbol{\Sigma}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}}^{-\mathsf{T}}$, or $(\mathbf{K}^{-\mathsf{T}}\boldsymbol{\Sigma}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}}^{-\mathsf{T}})^{\mathsf{T}}(\mathbf{K}^{-\mathsf{T}}\boldsymbol{\Sigma}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}}^{-\mathsf{T}})$. Experimentally we find the third option is actually the most stable and quickest in practice. We can then form $\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}}$ solve directly $\boldsymbol{\mu}_{\hat{\mathbf{f}}} = \boldsymbol{\Sigma}_{\widehat{\mathbf{f}}}\mathbf{b}$, where $\mathbf{b}$ is found first as $\mathbf{H}_1^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{g} - \boldsymbol{\Sigma}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}}^{-1}\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_0}\backslash(\mathbf{H}_0^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{g})$.

In the case that the noise level (and hence the regularisation) is low, and assuming there is sufficient memory to form $\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}}$, the fastest and most accurate solution seems to be found via Gaussian elimination. In most other cases, and when the observation region or boundary region are large enough, CG based solution will be more practical (we may need to run many iterations to achieve the same convergence in low-noise cases however).

The solution with CGLS does not follow exactly due to the negative factor in the covariance (that is, $\mathbf{M}^{\mathsf{T}}\mathbf{M}$ is not found trivially). Rather we use regular CG but with the matrix we are trying to invert partitioned similarly as $\check{\mathbf{M}}\mathbf{M}$, where $\boldsymbol{\Sigma}_{\mathbf{f}_1}^{-1} = \mathbf{C}_1^{\mathsf{T}}\mathbf{C}_1$:

$$\mathbf{M} = \begin{bmatrix} \sigma_w^{-1}\mathbf{H}_1 \\ \hline \mathbf{C}_1 \\ \hline \mathbf{K}^{-\mathsf{T}}\boldsymbol{\Sigma}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}}^{-\mathsf{T}} \end{bmatrix}, \qquad \text{and} \qquad \check{\mathbf{M}} = \begin{bmatrix} \sigma_w^{-1}\mathbf{H}_1 \\ \hline \mathbf{C}_1 \\ \hline -\mathbf{K}^{-\mathsf{T}}\boldsymbol{\Sigma}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}}^{-\mathsf{T}} \end{bmatrix}^{\mathsf{T}}. \tag{8.23}$$

To apply CG for medium sized images, e.g. up to about 100 pixels square with a 8 pixel wide boundary on each side, we can precompute the factor $(\mathbf{K}^{-\mathsf{T}}\boldsymbol{\Sigma}^{-\mathsf{T}}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}})$, and multiplication of a vector by this matrix is the only additional cost compared to the extended image case. As this matrix is not very sparse, the cost of multiplication is high compared to multiplying by $\mathbf{C}$, and the execution time of the CG iterations becomes slower. For very large images, this factor, which typically requires about half as much memory as the full matrix $\boldsymbol{\Sigma}_{\widehat{\mathbf{f}_0}}$ may become too large to store. In this case, the multiplication by a vector $\boldsymbol{x}$ is found as $\mathbf{K}^{\mathsf{T}}\backslash(\boldsymbol{\Sigma}^{-\mathsf{T}}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}}\boldsymbol{x})$. This is actually a lot slower as Gaussian elimination is performed at each CG iteration, but the whole algorithm then does not require any more memory than storing the original sparse matrices $\mathbf{H}_0$ and $\mathbf{C}$. In practice for larger images, it may be more sensible to partition them into blocks and solve each locally with marginalised boundaries, as suggested in [37].

8.6.3 Comparison to Calvetti & Somersalo's method

In [37], an alternative form of the linear system of the marginalised boundary model that may be solved using CGLS is found. The connections between their method and (8.13) are described in §E.3. The final form that they use however requires finding several other matrices from those used in (8.13). In particular, they need to find the (inverse) Cholesky factorisations of several Schur complements, including (see §E.3 for the definitions):

$$\widetilde{\boldsymbol{\Sigma}^{-1}_{\mathbf{f}_0}} = \boldsymbol{\Sigma}^{-1}_{\mathbf{f}_1} - \boldsymbol{\Sigma}^{-1}_{\mathbf{f}_1,\mathbf{f}_0}\boldsymbol{\Sigma}_{\mathbf{f}_0}\boldsymbol{\Sigma}^{-\mathsf{T}}_{\mathbf{f}_1,\mathbf{f}_0} \tag{8.24}$$

$$\mathbf{Q}_{\widehat{\mathcal{W}}} = \mathbf{Q}_{\mathcal{W}} + \mathbf{H}_0\boldsymbol{\Sigma}_{\mathbf{f}_0}\mathbf{H}_0^{\mathsf{T}} \tag{8.25}$$

$$\widetilde{\boldsymbol{\Sigma}^{-1}_{\mathbf{f}_1}} = \boldsymbol{\Sigma}^{-1}_{\mathbf{f}_0} - \boldsymbol{\Sigma}^{-1}_{\mathbf{f}_1,\mathbf{f}_0}\boldsymbol{\Sigma}_{\mathbf{f}_1}\boldsymbol{\Sigma}^{-\mathsf{T}}_{\mathbf{f}_1,\mathbf{f}_0} \tag{8.26}$$

The second two terms in particular take longer to calculate and factorise than finding $(\mathbf{K}^{-\mathsf{T}}\boldsymbol{\Sigma}^{-\mathsf{T}}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}})$ which is the only term needed for the iterative solution proposed in this chapter. Because there are several more linear systems to solve or inverses to find in succession, and each step introduces additional errors, it seems the resulting matrices are also less well conditioned than with the solution in this chapter. A direct solution like the one proposed here is also prohibitive using the form in [37], shown in (E.64), due to the need to invert the full covariance matrix of the prior.

The resulting CGLS procedure solves the linear system as before but using the matrix

$$\mathbf{M} = \begin{bmatrix} (\mathbf{Q}_{\widehat{\mathcal{W}}}^{-\frac{1}{2}})^{\mathsf{T}}(\mathbf{H}_1 + \mathbf{H}_0\widetilde{\boldsymbol{\Sigma}_{\mathrm{f}_1}^{-1}}^{-1}\boldsymbol{\Sigma}_{\mathrm{f}_1,\mathrm{f}_0}^{-1}\boldsymbol{\Sigma}_{\mathrm{f}_1\mathrm{f}_0}\widetilde{\boldsymbol{\Sigma}_{\mathrm{f}_0}^{-1}}) \\ \widetilde{\boldsymbol{\Sigma}_{\mathrm{f}_0}^{-1}}^{-\frac{1}{2}} \end{bmatrix}, \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} (\mathbf{Q}_{\widehat{\mathcal{W}}}^{\frac{1}{2}})^{\mathsf{T}}\backslash\mathbf{g} \\ \mathbf{0}_{\mathrm{L}} \end{bmatrix}. \tag{8.27}$$

In practice, Calvetti & Somersalo appear to suggest to precalculate two of the factors in the top block of $\mathbf{M}$: $\mathbf{E}_1 = \left(\widetilde{\boldsymbol{\Sigma}_{\mathrm{f}_1}^{-1}}^{\frac{1}{2}}\right)\backslash\left(\left(\widetilde{\boldsymbol{\Sigma}_{\mathrm{f}_1}^{-1}}^{\frac{1}{2}}\right)^{\mathsf{T}}\backslash\mathbf{H}_0^{\mathsf{T}}\right)$ and $\mathbf{E}_2 = \left(\boldsymbol{\Sigma}_{\mathrm{f}_1}^{-\frac{1}{2}}\right)\backslash\left(\left(\boldsymbol{\Sigma}_{\mathrm{f}_1}^{-\frac{1}{2}}\right)^{\mathsf{T}}\backslash\boldsymbol{\Sigma}_{\mathrm{f}_1,\mathrm{f}_0}^{-1}\right)$, such that it can be multiplied by a vector $\mathbf{x}$ by solving

$$\mathbf{Q}_{\widehat{\mathcal{W}}}^{\frac{1}{2}}\backslash\left(\mathbf{H}_1\mathbf{x} + \mathbf{E}_1^{\mathsf{T}}\mathbf{E}_2^{\mathsf{T}}\left(\left(\widetilde{\boldsymbol{\Sigma}_{\mathrm{f}_0}^{-1}}^{\frac{1}{2}}\right)^{\mathsf{T}}\widetilde{\boldsymbol{\Sigma}_{\mathrm{f}_0}^{-1}}^{\frac{1}{2}}\mathbf{x}\right)\right).$$

This is the way we use to implement their method, although there are various other factorisations and orders of multiplying the matrices which may be more or less well conditioned.

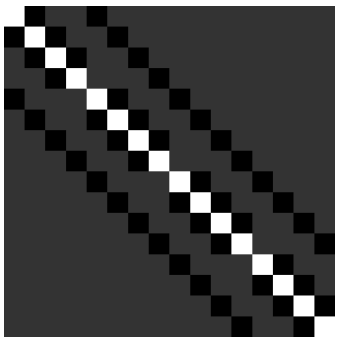
8.6.4 Modification of image prior

In general we will specify the factorisation of the image prior $\boldsymbol{\Sigma}_{\mathrm{f}_+}$ rather than the matrix itself — this simplifies the equations needed for CGLS. For instance, choosing $\mathbf{C}$ as representing the discrete Laplacian operator is equivalent to the stationary SAR prior. This is the prior used in the examples in this chapter.

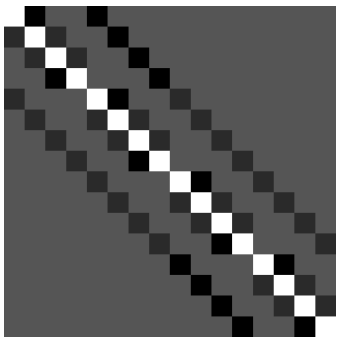
However, as noted in [37] and further discussed in [38], a standard BTTB representation of the Laplacian operator represents the Dirichlet boundary conditions on the extended image, i.e. that the image outside of $\mathcal{R}_+$ is zero. They suggest a modification to the matrix $\mathbf{C}$ such that the boundary pixels become uncorrelated with their interior/exterior neighbours, whilst the marginal autocovariance of each pixel remains approximately constant across the image.

We propose a simpler modification, as shown in fig. 8.4. That is simply to make each row of the matrix sum to the same value (zero), so that the correlation between a pixel and its available neighbours increases when there are fewer neighbours available. We adjust the weights in the matrix $\mathbf{C}$ so that interior pixels with four neighbours receive equal weight 0.25 as usual; boundary pixels with three neighbours receive a 0.5 weight on the interior pixel just inside the edge; and corner pixels receive weights 0.5 on both neighbours. In terms of an MRF representation, this amounts to increasing the potential for cliques that have fewer neighbours

inside the image. Observe that this modification is equivalent to evaluating the Laplacian under reflective boundary conditions, where the extra weight comes from the contribution the pixels just outside the boundary would give if they were there. This is a reasonable assumption for these pixels which are so far from the observed region that we have no actual observed information from them, and also results in maintaining a constant signal level towards the edge of the extended region. When we estimate more complicated priors as part of the image model, we will also use the modified Laplacian as an initialisation, to ensure homogeneity towards the boundary of the extended image.



(a) Before modification (Pixel values are black=-0.25, gray=0, white=1)



(b) After modification (Pixel values are black=-0.5, dark gray=-0.25, light gray=0, white=1)

Figure 8.4: *Modification of the* $\mathbf{C}$ *matrix using the discrete Laplacian. The example is shown for a* 4×4 *pixel image, before and after modification to give extra correlation to boundary pixels.*

8.6.5 Initialisation for extrapolation

Note to obtain the initial estimate for the CGLS algorithm in the extended boundary method, we perform extrapolation with no deblurring, i.e. run exactly the same CGLS procedure for a couple of hundred iterations, but replacing the matrix $\mathbf{H}_+$ with the identity, and using a large amount of regularisation ($\alpha = 10^{-2}$). The initialisation for this extrapolation is given by the observed image $\mathbf{g}$ placed onto the extended canvas, with the border set to be the mean value of $\mathbf{g}$. This initialisation helps give a good convergence of the actual restoration. The initialisations with and without the modified Laplacian are shown in fig. 8.8(a) and fig. 8.9(a).

8.7 Restoration examples

In fig. 8.5(a), we show an $\mathcal{R}_+ = 96 \times 96$ pixel region taken from the Lena image, which is blurred with a 3 pixel standard deviation Gaussian PSF with a support $\mathcal{S}_h$ of 16 pixels square. Noise is added to give a BSNR of 100dB — almost noise free, to show the limitations due to ill-conditioning of the method. This image is cropped to the $\mathcal{R} = 80 \times 80$ pixel observation region to give the observation shown in fig. 8.5(b). The initialisation for the extended boundary solution, after running 300 iterations of CGLS without deblurring is also shown in fig. 8.8(a).

The image is then deblurred with four different methods: the extrapolation method using CGLS; the two versions of the marginalised boundary solution proposed here (direct solution via Gaussian elimination and CG solution); and the CGLS marginalised boundary solution using the method from [37]. Fig. 8.6 and fig. 8.7 show the restorations using regularisation parameter $\alpha = 10^{-5}$ and 10^{-8}. All the CG solutions are shown after 1000 iterations — note that this has essentially converged in the $\alpha = 10^{-5}$ case but it can be seen that further improvement could be obtained with more iterations in the $\alpha = 10^{-8}$ case.

In table 8.1, the run times and memory requirements of the large matrices involved in each method are shown (assuming intermediate working matrices are deleted from memory when no longer needed). We have also included a row for a direct least squares (Gaussian elimination) implementation of the method from [37], solving the normal equations given by the matrix $\mathbf{M}$ in (8.27), (graphical result not shown). All the simulations were performed using MATLAB 7.4 under Linux on a 3GHz Intel Xeon quad core with 3GB of RAM (only 2 cores available to Matlab). Note also that the memory usage for the extrapolated boundary CGLS method is very low (only the storage for the Laplacian matrix $\mathbf{C}$). With priors using a larger support this will increase.

8.7.1 Unmodified prior for extended method

In fig. 8.9(a), the initialisation found when the standard Laplacian is used rather than the modified one described in §8.6.4 is used is shown. In fig. 8.9(b) and fig. 8.9(c) the resulting CGLS restorations are shown: the artefacts that result show that this modification is necessary to achieve satisfactory performance.

(a) Original image

(b) Blurred with 3 pixel Gaussian blur (16 pixel support) and cropped to $\mathcal{R}$

Figure 8.5: *Test images for comparing proposed boundary methods*

Method	Initialisation time (s)	Solution time (s)	Approx. memory required (MB)
Marginalisation — Direct (Gaussian Elimination)	139	27	320.7
Marginalisation — CG	22	130	91.0
Marginalisation — method from [37], CGLS	152	187	216.9
Marginalisation — method from [37], Direct	796	20	684.5
Extrapolation — CGLS	1.3	6.3	0.6

Table 8.1: *Run times and memory requirements for the different examples using the 96 pixel image*

(a) Direct solution (Gaussian elimination)

(b) CG solution

(c) CGLS solution using method from [37]

Figure 8.6: *Marginalised boundary restorations,* $\alpha = 10^{-5}$

(a) Direct solution (Gaussian elimination)

(b) CG solution

(c) CGLS solution using method from [37]

Figure 8.7: *Marginalised boundary restorations,* $\alpha = 10^{-8}$

(a) Initial estimate for extended image on $\Re_+$

(b) Restored with $\alpha = 10^{-5}$

(c) Restored with $\alpha = 10^{-8}$

Figure 8.8: *Solutions using extrapolation method and CGLS*

(a) Initial estimate

(b) Restored with $\alpha = 10^{-5}$

(c) Restored with $\alpha = 10^{-8}$

Figure 8.9: *Problems when standard Laplacian is used for extrapolation*

8.8 Discussion — Extrapolate or marginalise?

Although marginalisation is strictly more favourable than extrapolation, the computational demands are much higher, either using CG methods or via a direct solution. There are also likely to be more problems with ill-conditioning of the intermediate matrices. The results using the proposed methods appear to be slightly more efficient than those based on an implementation of the method in [37], but still an order of magnitude slower to initialise than the extrapolated CGLS solution which does not require calculation of any dense matrices and the bulk of its computation can be accelerated using the FFT.

Therefore in the remaining chapters of this work we will simply estimate the extended image and discard the border region,[4] which gives satisfactory results in practice, if we choose the prior carefully. For instance, if we are not too bothered about recovering the boundary, increasing the amount of smoothness in this region compared to the central region can help to avoid errors. When using an iterative BD method such as in Chapter 9, this may help to avoid oscillations in the boundary leading to instability of the solution.

In terms of notation, for simplicity in the remaining work we refrain from referring to $\mathbf{f}_+$ and $\mathbf{H}_+$ explicitly — the model in Chapter 6 is assumed to be placed across the full extended region $\mathcal{R}_+$, and any references to $\mathbf{f}$ or $\mathbf{H}$ or other variables involved in the image model refer to their extended versions.

8.8.1 Use of direct solution

Although the direct solution is slow to initialise and uses large amounts of memory, it does offer the best solution performance when regularisation is very low. Also given that it would take 5000 or even 10000 CG solutions to visually match the convergence in the $\alpha = 10^{-8}$ case, once the initialisation is done, the solution time is actually relatively fast. Observe also for a given PSF, the required matrices can be calculated and stored in advance, and then the direct solution found relatively quickly for multiple blurred regions. As the method deals correctly with the boundaries outside the observation window, a large image could be restored piecewise with each block independently calculated using this method, offering an efficient parallel implementation.

[4] We may also keep it if the extra pixels that are recovered are deemed useful — depending on the blur and noise levels, we can recover detail several extra pixels into the boundary

On the other hand, note that in practice 10^{-8} is a very small value of α, and with typical amounts of observation noise, fewer iteraitons would be sufficient. Ideally, in the low noise-case, preconditioning would help speed up the CG solution. A Fourier based preconditoner was tested, but with the full boundary model it caused problems with boundary artefacts where the method with no preconditioner was working. Further investigation would be required to find a preconditioner that works with the full boundary model.

8.8.2 Extensions for image inpainting

Another use for the methods described in this chapter is for more general image inpainting, for instance if there are missing, damaged or corrupted pixels in the blurred image. The methods are applied simply by redefining the observation region to be a non-rectangular region that masks out these pixels, and instead including them in $\mathcal{R}_0$ if they are within half the width of the PSF away from the closest observed pixel. If the work is extended to spatially varying blur due to depth of the scene, these methods can help recover missing data behind occlusions by masking different layers. In this way it could be useful, for instance, to see defocused objects through partial occlusions in front of the camera, such as a fence or branches of a tree.

Chapter 9
Variational Bayesian Approach

In this chapter, the VB method of distribution approximation will be considered for tackling the BD problem. This method has been used previously for BID, see e.g. [67, 132, 144, 153, 217]. The improvement over the method described in [153] that may be gained by incorporation of the new priors will be shown.

VB methods in general have been investigated in [14], and in [201] for applications in signal processing. Broadly speaking, it aims to allow inference to be performed in an approximate fashion, without resorting to more time consuming MCMC simulations, in situations when the posterior distribution is not tractable. It may also be viewed as a fully Bayesian version of the EM algorithm whereby both hidden variables and parameters are regarded as stochastic quantities from an unknown probability distribution.

The general method and derivations for calculating the approximate distributions needed for VB inference was presented in §5.2.5.1. Here we consider application of the method, using three different simplifications of the complete model presented in Chapter 6. We begin with a brief discussion of how to find the required distributions, and show an example to illustrate the application of the IVB procedure.

9.1 Distribution Approximations

The starting assumption we use in the VB method is that we can approximate the posterior $p(\Theta|\mathbf{g})$ by a product of separable distributions $q(\theta)$ over the variables of interest, θ, as shown in (5.18). Given this approximation, the distributions may be calculated using (5.25). We may then use the iterative VB approach to estimate the optimal distributions from the data; at each iteration, k, we denote these by $q^k(\theta)$.

The general procedure used in the following sections to find the distributions $q^k(\theta)$ for each θ is the same. We first evaluate (5.25) by substituting in required conditional PDFs; observe

that whilst in its general form, (5.25) contains the complete joint distribution, only terms dependent on the particular θ we are finding the distribution for need to be considered. In the tree structure of the graphical models used, these correspond to the conditional distribution of the "child" variable of θ, and the prior distribution of θ itself.

The next step is to perform the required expectations — again in these calculations we may ignore any terms independent of θ as they do not affect the functional form of $q^k(\theta)$. The expectations are taken with respect to the most recent distributional approximation of the other unknown variables present in the expression.

The final step is to recognise $q^k(\theta)$ as being in the form of a standard probability distribution. In general this may not always be possible and further approximations may have to be made to do so. However for the model presented here, due to the choice of conjugate priors we have used, the optimal form of each $q^k(\theta)$ turns out to be either a Gaussian or inverse Gamma distribution.

Intuitively, all we need to update the distribution $q^k(\theta)$ are its moments; since the the first two moments completely characterise the Gaussian and inverse Gamma distributions we are concerned with, these are all we need calculate to update the other distributions. We will use special notation to indicate these moments: $\mathbb{E}^k[\theta]$ is used to denote the expectation of θ with respect to the distribution $q^k(\theta)$, and $\mathrm{cov}^k[\theta]$ the covariance of θ at iteration k.

9.2 Simple VB example

In its most basic form, a simple blind restoration algorithm may be considered where the prior models for image, blur and noise are all assumed known, such that the unknown variables are $\boldsymbol{\Theta} = \{\mathbf{f}, \mathbf{h}\}$, as shown in fig. 9.1. The standard likelihood from the WGN model is used for $\mathbf{g}$, and the SAR model with fixed AR parameters given by the discrete Laplacian, $\mathbf{C}$ is used as the prior for $\mathbf{f}$ and $\mathbf{h}$, resulting in the distributions as shown in fig. 9.1.

Using (5.25) and letting $\theta = \mathbf{f}$, the approximate distribution for $\mathbf{f}$ is found as follows:

$$\log \widehat{q}(\mathbf{f}) \propto \mathbb{E}\left[\log\left(p(\mathbf{g}\,|\,\mathbf{f},\mathbf{h})\,p(\mathbf{f})\right)\right]_{q(\boldsymbol{\Theta}_{\setminus \mathbf{f}})} \tag{9.1}$$

$$\propto \mathbb{E}\left[\log\left(\sigma_w^{-mn}\exp\left[\sigma_w^{-2}\|\mathbf{g}-\mathbf{H}\mathbf{f}\|^2\right]\cdot\sigma_v^{-mn}\exp\left[\sigma_v^{-2}\|\mathbf{C}\mathbf{f}\|^2\right]\right)\right]_{q(\boldsymbol{\Theta}_{\setminus \mathbf{f}})} \tag{9.2}$$

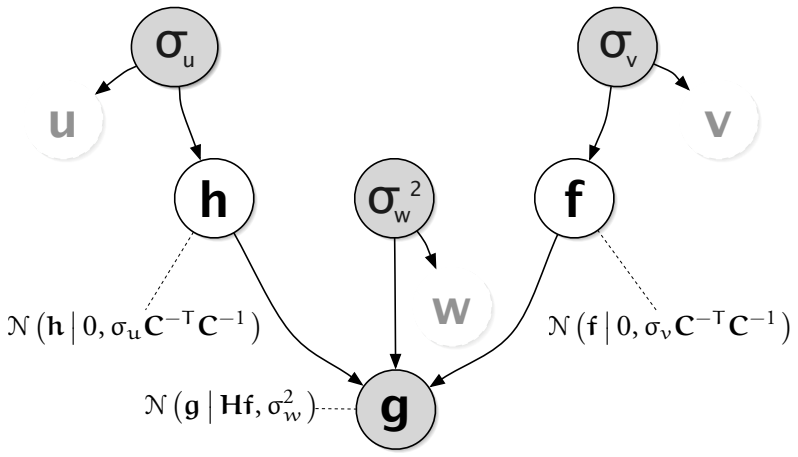


Figure 9.1: *Graphical model for the simple stationary model, with* $\boldsymbol{\Theta} = \{\mathbf{f}, \mathbf{h}\}$

$$\propto -\frac{1}{2}\left(\mathbb{E}\left[\sigma_w^{-2}\|\mathbf{g}-\mathbf{H}\mathbf{f}\|^2\right]_{\mathrm{q}(\mathbf{h})} + \mathbb{E}\left[\sigma_v^{-2}\|\mathbf{C}\mathbf{f}\|^2\right]_{\mathrm{q}(\mathbf{h})}\right). \tag{9.3}$$

In (9.3), only the terms that depend on $\mathbf{f}$ need to be kept, as discussed in the previous section. These are the conditional distribution of the child variable $\mathbf{g}$ of $\mathbf{f}$ (which is the likelihood), and the prior for $\mathbf{f}$. Furthermore, terms not conditional on the other variables ($\mathbf{h}$ in this case) may be taken outside the expectations.

The next step is to identify a functional form for the distribution. We could choose to assign a particular distribution and optimise with respect to its parameters here; however, with some careful examination it should be seen that $\mathrm{q}(\mathbf{f})$ is already in the form of a Gaussian. This should be expected, due to the choice of a conjugate prior for $\mathbf{f}$ as the mean of a Gaussian likelihood. Hence the Gaussian distribution are the optimal form for the distribution $\mathrm{q}(\mathbf{f})$ in this case. Expanding the first expectation gives

$$\mathbb{E}\left[\sigma_w^{-2}\|\mathbf{g}-\mathbf{H}\mathbf{f}\|^2\right]_{\mathrm{q}(\mathbf{h})} = \sigma_w^{-2}\left(\mathbf{g}^{\mathsf{T}}\mathbf{g} - 2\mathbf{g}^{\mathsf{T}}\mathbb{E}\left[\mathbf{H}\right]_{\mathrm{q}(\mathbf{h})}\mathbf{f} + \mathbf{f}^{\mathsf{T}}\mathbb{E}\left[\mathbf{H}^{\mathsf{T}}\mathbf{H}\right]_{\mathrm{q}(\mathbf{h})}\mathbf{f}\right). \tag{9.4}$$

Now with the deduction that $\widehat{\mathrm{q}}(\mathbf{f})$ is Gaussian, that is

$$\widehat{\mathrm{q}}(\mathbf{f}) \sim \mathcal{N}\left(\mathbf{f}\,\middle|\,\boldsymbol{\mu}_{\widehat{\mathrm{q}}(\mathbf{f})}, \boldsymbol{\Sigma}_{\widehat{\mathrm{q}}(\mathbf{f})}\right) \tag{9.5}$$

$$\propto -\frac{1}{2}\left(\mathbf{f}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}_{\widehat{\mathrm{q}}(\mathbf{f})}\mathbf{f} - 2\mathbf{f}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}_{\widehat{\mathrm{q}}(\mathbf{f})}\boldsymbol{\mu}_{\widehat{\mathrm{q}}(\mathbf{f})} + \boldsymbol{\mu}^{\mathsf{T}}_{\widehat{\mathrm{q}}(\mathbf{f})}\boldsymbol{\Sigma}^{-1}_{\widehat{\mathrm{q}}(\mathbf{f})}\boldsymbol{\mu}_{\widehat{\mathrm{q}}(\mathbf{f})}\right), \tag{9.6}$$

equating quadratic and linear terms in $\mathbf{f}$ gives

$$\boldsymbol{\Sigma}_{\widehat{q}(\mathbf{f})} = \left(\sigma_w^{-2}\mathbb{E}\left[\mathbf{H}^T\mathbf{H}\right]_{q(\mathbf{h})} + \sigma_v^{-2}\mathbf{C}^T\mathbf{C}\right)^{-1} \tag{9.7}$$

$$\boldsymbol{\mu}_{\widehat{q}(\mathbf{f})} = \boldsymbol{\Sigma}_{\widehat{q}(\mathbf{f})}\left(\sigma_w^{-2}\mathbb{E}\left[\mathbf{H}\right]_{q(\mathbf{h})}^T \mathbf{g}\right). \tag{9.8}$$

The derivation for $\widehat{q}(\mathbf{h})$ arrives at a similar result due to the symmetry of the model. If the IVB procedure is now used, the algorithm will alternate between updating $\widehat{q}(\mathbf{f})$ and $\widehat{q}(\mathbf{h})$. To indicate the distributions at a particular iteration, k, the notation $q^k(\boldsymbol{\theta})$ will be introduced.

Adopting this notation, the IVB procedure amounts to updating the moments of each distribution based on the expectations of the other previously found $q(\cdot)$ distributions; that is the expectations are taken with respect to the most recent distributional approximation of the other unknown variables present in the expression.

Returning to the example, equations (9.7) and (9.8) will provide the necessary moments updating $q^k(\mathbf{f})$, however firstly the expectation of $\mathbf{H}^T\mathbf{H}$ needs to be defined. The notation $\text{cov}^k\left[\mathbf{X}^T\right] \triangleq \mathbb{E}^k\left[\mathbf{X}^T\mathbf{X} - \mathbb{E}^k[\mathbf{X}]^T\mathbb{E}^k[\mathbf{X}]\right]$ representing the covariance of a Toeplitz matrix will be introduced to allow this term to be expanded. Further details of this notation and calculating its entries are given in Appendix G, but it is sufficient to note that it can be calculated based upon the elements of $\text{cov}^k[\mathbf{x}]$, where $\mathbf{x}$ is the defining sequence of the convolution. Therefore the term in question may be expanded as

$$\mathbb{E}\left[\mathbf{H}^T\mathbf{H}\right]_{q^{k-1}(\mathbf{h})} = \text{cov}^{k-1}\left[\mathbf{H}^T\right] + \mathbb{E}^k[\mathbf{H}]^T\mathbb{E}^k[\mathbf{H}] \tag{9.9}$$

Now equations (9.7) and (9.8) may be rewritten at iteration k as

$$\text{cov}^k[\mathbf{f}] = \left(\sigma_w^{-2}\left(\text{cov}^{k-1}\left[\mathbf{H}^T\right] + \mathbb{E}^{k-1}[\mathbf{H}]^T\mathbb{E}^{k-1}[\mathbf{H}]\right) + \sigma_v^{-2}\mathbf{C}^T\mathbf{C}\right)^{-1} \tag{9.10}$$

$$\mathbb{E}^k[\mathbf{f}] = \text{cov}^k[\mathbf{f}]\left(\sigma_w^{-2}\mathbb{E}^{k-1}[\mathbf{H}]^T\mathbf{g}\right). \tag{9.11}$$

Similarly, the update equations for $q^k(\mathbf{h})$ are found as

$$\text{cov}^k[\mathbf{h}] = \left(\sigma_w^{-2}\left(\text{cov}^k\left[\mathbf{F}^T\right] + \mathbb{E}^k[\mathbf{F}]^T\mathbb{E}^k[\mathbf{F}]\right) + \sigma_u^{-1}\mathbf{C}^T\mathbf{C}\right)^{-1} \tag{9.12}$$

$$\mathbb{E}^k[\mathbf{h}] = \text{cov}^k[\mathbf{h}]\left(\sigma_w^{-2}\mathbb{E}^k[\mathbf{F}]^T\mathbf{g}\right). \tag{9.13}$$

In practice, calculating the terms $\text{cov}^{\mathsf{k}}\left[\mathbf{H}^{\mathsf{T}}\right]$ and $\text{cov}^{\mathsf{k}}\left[\mathbf{F}^{\mathsf{T}}\right]$ is prohibitive for larger image sizes, as they require finding and summing over the covariance matrices $\text{cov}^{\mathsf{k}}\left[\mathbf{h}\right]$ and $\text{cov}^{\mathsf{k}}\left[\mathbf{f}\right]$.

When solving the equations in the Fourier domain with assumed circulant boundaries, the terms simply become scaled versions of the corresponding covariance matrix, as described in Appendix G, which makes the update equations simple to implement. If we wish to calculate them using the proper extended boundary conditions, this unfortunately implies inversion of the covariance matrices as an initial step, which adds prohibitively to the cost of the algorithm. As we are actually more interested in the expected values of the variables rather than their covariances, the equations can normally be solved efficiently using CGLS, omitting the $\text{cov}^{\mathsf{k}}\left[\mathbf{H}^{\mathsf{T}}\right]$ and $\text{cov}^{\mathsf{k}}\left[\mathbf{F}^{\mathsf{T}}\right]$ terms. We could attempt to include them iteratively, performing an inner CGLS step in a similar manner to the iterative solution proposed to the marginalised boundary problem in §8.6. However, this would still be costly and we have found it simpler to approximate these terms as zero, resulting in degenerate distributions (delta functions placed at the mean).

As it turns out, the magnitude of the maximum coefficients in the estimated $\text{cov}^{\mathsf{k}}\left[\mathbf{H}^{\mathsf{T}}\right]$ are at least two orders of magnitude smaller than the smallest errors we encounter in practice in the PSF (and hence in $\mathbb{E}\left[\mathbf{H}\right]$), and similarly for the image. So the terms appear not to have any significant effect on the restoration, as is confirmed when we remove them in the Fourier domain estimation. When estimating scalar variance parameters in the following sections, they may have more of an effect, as the values are summed over the support of the matrices' defining sequences. Therefore one option would be to approximate these matrices by their circulant counterparts. Although we show the full derivations in Appendix F, for simplicity in our implementation however we have treated them as zero. Note that this results in an AM type algorithm — in this simple case where only $\mathbf{h}$ and $\mathbf{f}$ are estimated, the resulting algorithm is equivalent to that used in [235].

9.2.1 VB discussion

It will be seen that the prime advantages of the VB approach are a straightforward and standardised method of producing an iterative solution to a complicated hierarchical Bayesian network model, which allows for quantifying and making use of the uncertainty in the estimates of each variable. On the other hand, the disadvantages include appropriateness of the separable approximation, and the fact that the covariances (or other moments) of the resulting

approximate distributions may be hard to evaluate in practice, due to their large size, as is the case for image and blur distributions. In this case the higher order moments may be chosen to be ignored, meaning that deterministic updates are used; we may describe this in a Bayesian sense by assuming each $q(\cdot)$ is a degenerate distribution, that is a delta function at the previous expected value. This results in an algorithm that alternates between finding the conditional expected value of each variable in turn, which for Gaussian distributions is the same as the maximising the exponent, hence an AM or ICM type algorithm results, or an EM algorithm if some of the variables are regarded as stochastic (the hidden variables) and others as degenerate (the unknown parameters).

Furthermore, in certain cases we may find that the resulting $q(\cdot)$ distributions are not in a standard form, which means that the moments cannot be found, without making further approximations. For example, multivariate distributions may be further split up into independent variables, or a generalised-VB approach may be used as described previously. Another possibility is to perform the required intractable expectations using MCMC, resulting in hybrid approaches, but the usefulness of this depends on the trade off between speed and performance versus going to a full MCMC approach.

In the remainder of the chapter, the VB approach described above will be applied to the image models that are under consideration, starting with a simple model and building up to the full non-stationary image/blur model.

9.3 VB applied to stationary image model

In this section we will begin by examining the application of the IVB algorithm to a stationary image model. This model was used by Molina *et al.*in [153], and will be used here as a baseline for comparisons with the extended non-stationary versions of the model which will be presented. The method is described in this section, with experiments shown as a comparison in the following section.

This model introduces the hyperpriors on the variance parameters, with the distributions shown in fig. 9.2. The distributions for $\mathbf{f}$ and $\mathbf{h}$ are as given in the previous section, except we use the expected values for the variances. The distributions for the scalar variance parameters are found as in the following example.

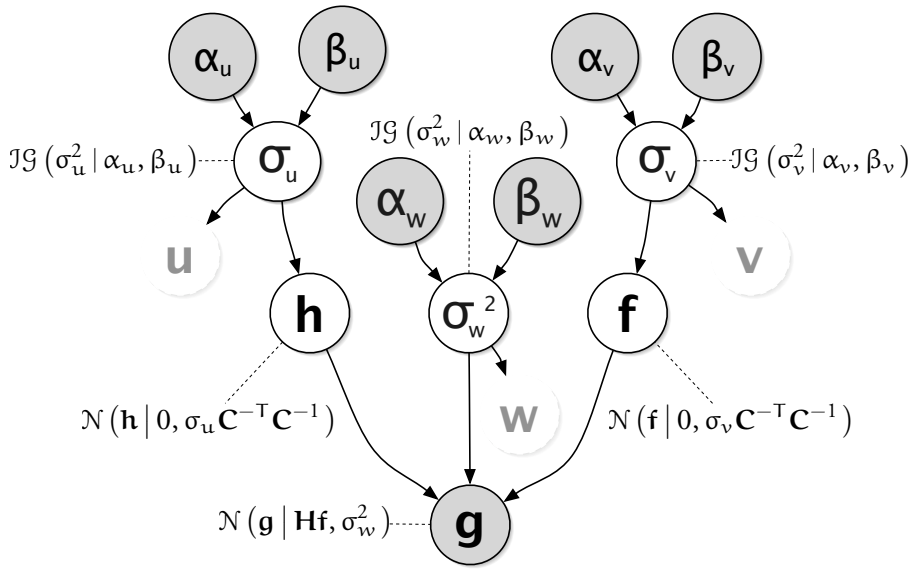


Figure 9.2: *Graphical model for the stationary model*

For the distribution $\widehat{q}(\sigma_v)$, using (5.25) (again ignoring terms independent of σ_v), we find:

$$\log \widehat{q}(\sigma_v) \propto \mathbb{E}\left[\log\left(p(\mathbf{f} \mid \sigma_v)\, p(\sigma_v)\right)\right]_{q(\boldsymbol{\Theta}_{\setminus\sigma_v})} \tag{9.14}$$

$$\propto \log(p(\sigma_v)) - \frac{1}{2} L_+ \log(\sigma_v^2) + \mathbb{E}\left[\sigma_{v_r}^{-2} \boldsymbol{v}^{\mathsf{T}}\boldsymbol{v}\right]_{q(\mathbf{f})} \tag{9.15}$$

$$\Rightarrow \widehat{q}(\sigma_v) = \mathcal{IG}\left(\sigma_v^2 \,\middle|\, \alpha_v + \frac{L_+}{2}, \beta_v + \frac{1}{2}\mathbb{E}\left[\boldsymbol{v}^{\mathsf{T}}\boldsymbol{v}\right]_{q(\mathbf{f})}\right) \tag{9.16}$$

We have identified this as being in the form of another inverse Gamma distribution. The other variance distributions take similar forms.[1] We then use the parameterisation from (6.48) to define the mean of the distribution at iteration k as

$$\mathbb{E}^k\left[\sigma_v^2\right] = \gamma_{\sigma_v^2}\mathbb{E}^0\left[\sigma_v^2\right] + (1-\gamma_{\sigma_v^2})\frac{1}{L_+}\mathbb{E}\left[\|\boldsymbol{v}^{\mathsf{T}}\boldsymbol{v}\|^2\right]_{q^{k-1}(\mathbf{f})} \tag{9.17}$$

where $\mathbb{E}^0\left[\sigma_v^2\right]$ is the mean of the prior $p(\sigma_v)$, and $\gamma_{\sigma_v^2}$ is the confidence parameter with a value between zero and one, with values near one enforcing the prior completely and zero causing the parameter to be estimated solely from the data (as an ML estimate). All that is left to find[2]

[1] Observe that these variance distributions are all independent, even without the separable approximation, i.e. $q(\sigma_v, \sigma_u, \sigma_w) = q(\sigma_v)\, q(\sigma_u)\, q(\sigma_w)$ in any case, as they do not share any child nodes in the tree.

[2] Note that we do not need to calculate the variances of the $q(\cdot)$ distributions for the variance parameters, because none of the other distributions depend on them.

is the expectation $\mathbb{E}\left[\|\boldsymbol{\nu}^{\mathsf{T}}\boldsymbol{\nu}\|^2\right]_{q^{k-1}(\mathbf{f})}$. We have that

$$\mathbb{E}\left[\boldsymbol{\nu}^{\mathsf{T}}\boldsymbol{\nu}\right]_{q^k(\mathbf{f})} = \mathbb{E}\left[\|\mathbf{C}_f\mathbf{f}\|^2\right]_{q^k(\mathbf{f})} \tag{9.18}$$

$$= \mathbb{E}^k[\boldsymbol{\nu}]^{\mathsf{T}}\mathbb{E}^k[\boldsymbol{\nu}] + \mathrm{tr}\left[\mathbf{C}_f^{\mathsf{T}}\mathbf{C}_f\,\mathrm{cov}^k[\mathbf{f}]\right]. \tag{9.19}$$

Note that (9.19) is derived using the identities $\mathbf{x}^{\mathsf{T}}\mathbf{x} = \mathrm{tr}\left[\mathbf{x}\mathbf{x}^{\mathsf{T}}\right]$ and $\mathbb{E}\left[\mathbf{x}\mathbf{x}^{\mathsf{T}}\right] = \mathbb{E}[\mathbf{x}]\,\mathbb{E}[\mathbf{x}]^{\mathsf{T}} + \mathrm{cov}[\mathbf{x}]$. Again evaluation of the final term inside the trace is straightforward in the Fourier domain, however we have ignored it in our simulations for simplicity (it could also be approximated by a BCCB matrix). The result for σ_u is identical in form, and the full derivation of the terms needed to find the expectation $\mathbb{E}\left[\|\mathbf{g}-\mathbf{H}\mathbf{f}\|^2\right]_{q^{k-1}(\mathbf{h},\mathbf{f})}$ used in the distribution $q(\sigma_w)$ is given in Appendix F. formulation with the extrapolated boundary conditions

9.4 Nonstationary Variance Algorithm

We now consider a basic extension to the model from the previous section, by allowing the variances $\boldsymbol{\sigma}_\nu = [\sigma^2_{\nu_1}\cdots\sigma^2_{\nu_R}]^{\mathsf{T}}$ in the image to adapt locally within each region of the image. We will describe the full algorithm before presenting results.

We use the prior

$$p(\mathbf{f}\,|\,\boldsymbol{\sigma}_\nu) \propto \prod_{r=1}^{R}\sigma_{\nu_r}^{-L_r}\exp\left[-\frac{1}{2}\sigma_{\nu_r}^{-2}\boldsymbol{\nu}_r^{\mathsf{T}}\boldsymbol{\nu}_r\right] \tag{9.20}$$

where $\boldsymbol{\nu}_r = \mathbf{C}_{r_+}\mathbf{f}_{r_+}$, which is a simplification of the full AR model — using a fixed covariance $\mathbf{C}_{r_+}$ given by the discrete Laplacian — but it allows for local variance parameters.

Similar to the result for the stationary variance case, we find the approximate distribution

$$\log\widehat{q}(\boldsymbol{\sigma}_\nu) \propto \log(p(\boldsymbol{\sigma}_\nu)) - \frac{1}{2}\sum_{r=1}^{R}L_r\log(\sigma^2_{\nu_r}) + \mathbb{E}\left[\sigma_{\nu_r}^{-2}\boldsymbol{\nu}_r^{\mathsf{T}}\boldsymbol{\nu}_r\right]_{q(\mathbf{f})} \tag{9.21}$$

$$\Rightarrow \widehat{q}(\boldsymbol{\sigma}_\nu) = \prod_{r=1}^{R}\mathcal{IG}\left(\sigma^2_{\nu_r}\,\middle|\,\alpha_{\nu_r}+\frac{L_r}{2},\beta_{\nu_r}+\frac{1}{2}\mathbb{E}\left[\boldsymbol{\nu}_r^{\mathsf{T}}\boldsymbol{\nu}_r\right]_{q(\mathbf{f})}\right) \tag{9.22}$$

This is in the form of a product of independent inverse Gamma distributions, over the variance of each region. The other distributions take the same forms as before, except we use the expectation of the covariance matrix $\mathbf{Q}_\nu$ (parameterised by $\boldsymbol{\sigma}_\nu$) in place of the scalar image variance

from before. Summarising the moments we need to calculate to represent the distributions, we end up with Algorithm 1.

In practice, to solve in the spatial domain, we approximate $q(\mathbf{f})$ and $q(\mathbf{h})$ as degenerate distributions and do not calculate the terms $\text{cov}^k\left[\mathbf{F}^T\right]$ and $\text{cov}^k\left[\mathbf{H}^T\right]$. Instead the solutions for $\mathbf{f}$ and $\mathbf{h}$ are found using CGLS as described in Chapter 8 for the extended image.

Input: $\mathbf{g}, \mathbb{E}^0\left[\sigma_u^2\right], \mathbb{E}^0\left[\sigma_w^2\right], \mathbb{E}^0\left[\sigma_{v_r}^2\right], \gamma_{\sigma_u^2}, \gamma_{\sigma_w^2}, \gamma_{\sigma_{v_r}^2}, \rho_0$

Output: $\hat{\mathbf{f}} = \mathbb{E}^k[\mathbf{f}], \hat{\mathbf{h}} = \mathbb{E}^k[\mathbf{h}]$

$\mathbb{E}^0[\mathbf{f}] = [\mathbf{g}^T, \bar{g}\mathbf{1}]^T$ (Where $\bar{g}$ is the mean of $\mathbf{g}$) ;

$\mathbb{E}^0[\mathbf{h}] = \text{GaussianPSF}(\rho_0)$;

$k = 1$;

while $\left(\|\mathbb{E}^k[\mathbf{f}]\|^2 - \|\mathbb{E}^{k-1}[\mathbf{f}]\|^2\right) < \epsilon\|\mathbb{E}^{k-1}[\mathbf{f}]\|^2$ **do**

$\mathbb{E}^k\left[\sigma_u^2\right] = \gamma_{\sigma_u^2}\mathbb{E}^0\left[\sigma_u^2\right] + \frac{(1-\gamma_{\sigma_u^2})}{m_h n_h}\mathbb{E}\left[\|\mathbf{C}_h\mathbf{h}\|^2\right]_{q^{k-1}(\mathbf{h})}$;

$\mathbb{E}^k\left[\sigma_w^2\right] = \gamma_{\sigma_w^2}\mathbb{E}^0\left[\sigma_w^2\right] + (1-\gamma_{\sigma_w^2})\frac{1}{mn}\mathbb{E}\left[\|\mathbf{g}-\mathbf{H}\mathbf{f}\|^2\right]_{q^{k-1}(\mathbf{h})q^{k-1}(\mathbf{f})}$;

for $r = 1 \cdots R$: **do**

$\mathbb{E}^k\left[\sigma_{v_r}^2\right] = \gamma_{\sigma_{v_r}^2}\mathbb{E}^0\left[\sigma_{v_r}^2\right] + (1-\gamma_{\sigma_{v_r}^2})\frac{1}{L_r}\mathbb{E}\left[\|\mathbf{C}_{\mathbf{f}_{r_+}}\mathbf{f}_{r_+}\|^2\right]_{q^{k-1}(\mathbf{f})}$, $\forall r$;

$\left(\text{cov}^k[\mathbf{f}]\right)^{-1} = \mathbf{C}_f^T\mathbb{E}^k[\mathbf{Q}_v]^{-1}\mathbf{C}_f + \mathbb{E}^k\left[\sigma_w^2\right]^{-1}\left(\mathbb{E}^{k-1}[\mathbf{H}]^T\mathbb{E}^{k-1}[\mathbf{H}] + \text{cov}^{k-1}\left[\mathbf{H}^T\right]\right)$;

$\mathbb{E}^k[\mathbf{f}] = \mathbb{E}^k\left[\sigma_w^2\right]^{-1}\text{cov}^k[\mathbf{f}]\,\mathbb{E}^{k-1}[\mathbf{H}]^T\mathbf{g}$;

$\left(\text{cov}^k[\mathbf{h}]\right)^{-1} = \mathbb{E}^k\left[\sigma_u^2\right]^{-1}\mathbf{C}_h^T\mathbf{C}_h + \mathbb{E}^k\left[\sigma_w^2\right]^{-1}\left(\mathbb{E}^k[\mathbf{F}]^T\mathbb{E}^k[\mathbf{F}] + \text{cov}^k\left[\mathbf{F}^T\right]\right)$;

$\mathbb{E}^k[\mathbf{h}] = \mathbb{E}^k\left[\sigma_w^{d'2}\right]^{-1}\text{cov}^k[\mathbf{h}]\,\mathbb{E}^k[\mathbf{F}]^T\mathbf{g}$;

$k = k + 1$;

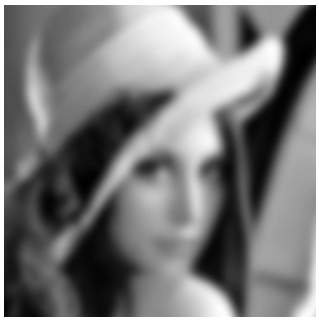
Algorithm 1: *Iterative VB for Nonstationary variance BID.*

9.4.1 Examples

We have implemented the two methods describe above: the first is similar to that in [153], the second method we presented in the paper [26]. We present the following results for comparison with the other methods we use.

In fig. 9.3, we see a comparison between restorations using the stationary model in the spatial domain and the Fourier domain, for a 3 pixel Gaussian blurred image at 40dB BSNR. Each is blurred using the same boundary conditions used in the restoration model, i.e. extended and circulant respectively (if we use the extended boundaries for the Fourier restoration we end up with ringing artefacts as described in Chapter 8.

Part of the reason for the better quality of the spatial domain restorations (apart from the extended boundary) is that we can use a smaller estimated PSF support — we have used 39×39 pixels in the spatial domain case, more than large enough for the coefficients of the Gaussian PSF to drop below the noise floor.[3] The Fourier solution requires the PSF support to be the same size as the image, a major disadvantage in practice.



(a) Observed image

(b) Restored via CGLS — ISNR=2.66dB, SSIM=0.695

(c) Observed (Circulant boundary)

(d) Restored in Fourier domain (using the method described in [153]) — ISNR=2.27dB, SSIM=0.657

Figure 9.3: *Restorations with the stationary model — 40dB BSNR,* $\mathbb{E}^0\left[\sigma_u^2\right] = 10^{-8}$

In fig. 9.4(a) we show a cross-section through the centre of the PSFs recovered by each method, compared with the true PSF. Recall that due to the scale ambiguity, any scaled version of the PSF gives the same result, so in order not to distort the errors, they should really be considered using optimally scaled versions of the PSFs. In fig. 9.4(b) we show the error of the estimated

[3] we could have used an even smaller support in practice without really affecting the results — in fact they would most likely be marginally better due to pruning of noise where the PSF should be zero — but using a larger support size than necessary shows that the proposed algorithms are not too sensitive to the size we use, so long as it is greater or equal to the true support.

PSFs found after scaling each such that the maximum value is the same as the true PSF.

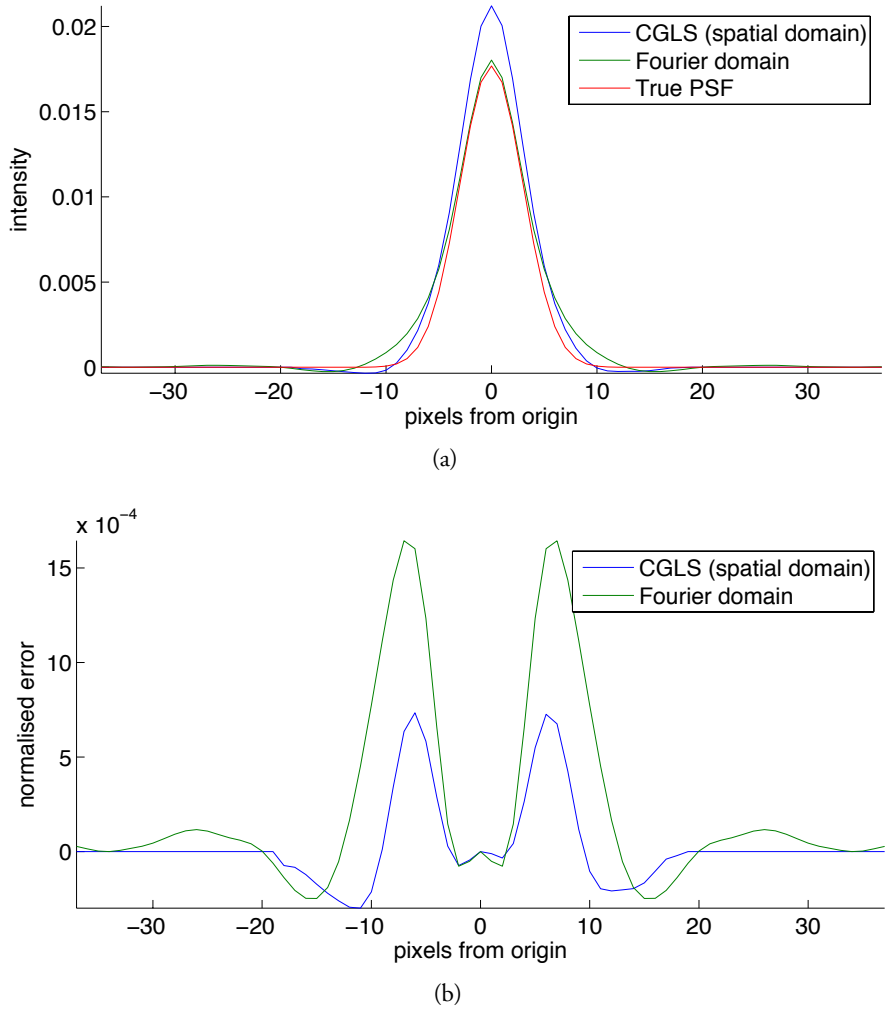


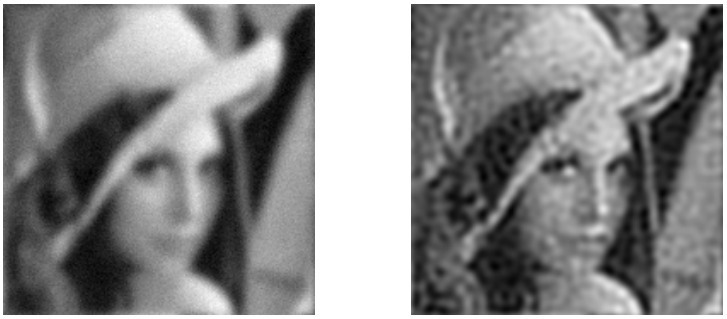
Figure 9.4: *(a) PSFs and (b) their errors, estimated with the stationary model — 40dB BSNR,* $\mathbb{E}^0\left[\sigma_u^2\right] = 10^{-8}$

In fig. 9.5, we show similar results for the 20dB BSNR case; note there is less improvement in the spatial domain this time as noise amplification errors dominates the error in the solution, rather than errors in the PSF (these are shown in fig. 9.6).

In both sets of simulations we set the (hyper)-hyperparameters as $\mathbb{E}^0\left[\sigma_u^2\right] = 10^{-8}; \mathbb{E}^0\left[\sigma_w^2\right] = 0.5; \mathbb{E}^0\left[\sigma_{v_r}^2\right] = 400, r \in \mathcal{R}_0; \gamma_{\sigma_u^2} = 0.5; \gamma_{\sigma_w^2} = 0.5; \gamma_{\sigma_{v_r}^2} = 0.1; \rho_0 = 1.0$.

(a) Observed image

(b) Restored via CGLS — ISNR=1.70dB, SSIM=0.602



(c) Observed (Circulant boundary)

(d) Restored in Fourier domain (using the method described in [153]) — ISNR=1.38dB, SSIM=0.568

Figure 9.5: *Restorations with the stationary model — 20dB BSNR,* $\mathbb{E}^0\left[\sigma_u^2\right] = 10^{-8}$

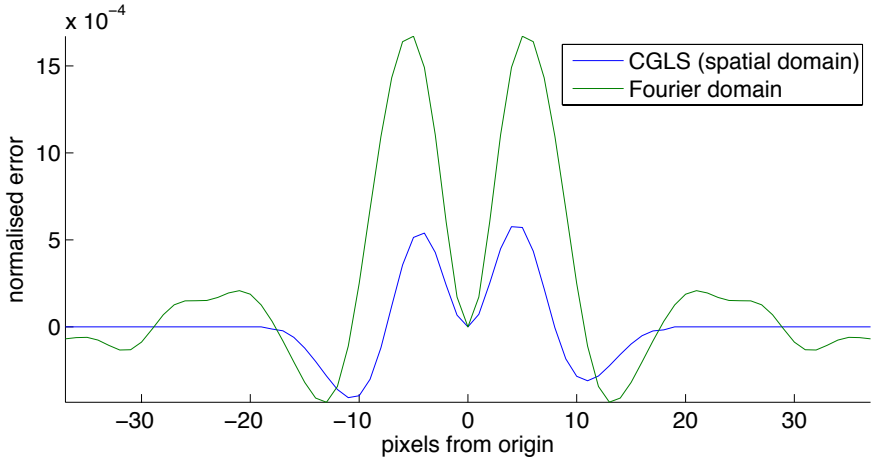


Figure 9.6: *PSF errors, estimated with the stationary model — 20dB BSNR,* $\mathbb{E}^0\left[\sigma_u^2\right] = 10^{-8}$

(a) Restored via CGLS — ISNR=2.78dB, SSIM=0.685

(b) Restored in Fourier domain — ISNR=0.98dB, SSIM=0.590

Figure 9.7: *Restorations with the stationary model — 40dB BSNR,* $\mathbb{E}^0\left[\sigma_u^2\right] = 10^{-9}$ *(to show effect of adjusting hyperparameters)*

9.4.1.1 Discussion: effect of parameter adjustments

Despite the hierarchical framework, there is still an important role played by setting the "known" hyper-hyperparameters correctly. Furthermore, there are other parameters in the implementation that are not strictly part of the model but will affect the numerical conditioning and convergence of the algorithm. Going into detail of all the possibilities would be beyond the scope of this thesis, but some general observations are made on their effects, based on running a large number of experiments.

One significant parameter that has a major role on convergence is the number of CG steps applied per VB iteration. With the Fourier solution we converge fully for $\mathbf{f}$ and $\mathbf{h}$ at each VB iteration. However with the CG based spatial domain estimates, we have an initial guess available: at each VB iteration k we can initialise with the previous estimate at iteration $k-1$ and just take a few CG steps. This not only speeds up the algorithm but also means the updates are only in a direction relevant to the solution.

Thus it is typical to set a tolerance on the residual norm, giving a maximum number of iterations before early termination of CGLS. This avoids "over-converging" based on noise and discretisation errors in the system, and acts as a form of regularisation. When we run many iterations nested in the VB iterations, even the machine tolerance has an effect. Therefore we choose to terminate early. Typically for the problems we have considered, we have found that setting the tolerances per pixel of the residual norm to the order of 5×10^{-5} for the

image and 25 for the PSF provides the optimal trade off between speed and accuracy. There is another reason: if we consider the alternating minimisation over $\mathbf{f}$ and $\mathbf{h}$ as part of a joint non-linear minimisation, then taking smaller steps alternately should be closer to the simultaneous minimisation than using larger steps. Over the course of all the VB iterations we run enough CG steps to converge to the correct solution.

In conjunction with these parameters, the prior for the blur has the most impact on the speed of convergence of the algorithm, because typically the initial guess for the PSF is a long way from the solution, whereas the error in the image is smaller. Because we have not calculated the full covariances in the expectations in practice, there is a tendency to underestimate the values of parameters such as σ_u^2. Thus the value found in practice is quite dependant on the prior mean. A smaller value of $\mathbb{E}^0\left[\sigma_u^2\right]$ tends to smooth the blur more rapidly; hence the solution is found in fewer VB iterations but it is generally of lower quality (e.g. more ringing or halo artefacts), as more CG iterations are applied to $\mathbf{f}$ and $\mathbf{h}$ alternately to perform these larger steps. An example of the effect of decreasing the prior expected value of σ_u^2 away from the value used in the other experiments is shown in fig. 9.7.

It should be noted that often changing one of these parameters has a moderate effect on the ISNR figure, the visual appearance is generally not significantly changed: as long as we pick a figure in the right ballpark the algorithm still converges, just it may take fewer or more iterations. Furthermore, when we move to the models using the more adaptive priors, it is also possible to use larger steps via more CG iterations, as the errors introduced by the larger steps are ameliorated by the stronger priors which better preserve the underlying image content.

9.4.2 Nonstationary examples: Problems with feedback and "Hot Pixels"

We show some results using the nonstationary variance algorithm, using different sized blocks, in fig. 9.8. Also shown are the errors in the corresponding PSF estimates in fig. 9.9. It appears there are some small sub-pixel shifts in the results, as can be seen in the asymmetry of the PSFs. Note that the PSF errors in the Fourier restorations are generally symmetric (and the solutions are not shifted). In order to evaluate the errors more precisely when there is a shift in the solution due to the shift ambiguity, we should also compensate for the misalignment.

Returning to the nonstationary model, we would like to make the variances as non-stationary as possible. This is subject to the local activity, such that smooth areas end up with small

(a) Blocksize = 1 — ISNR=2.62dB, SSIM=0.693

(b) Blocksize = 2 — ISNR=2.68dB, SSIM=0.696

(c) Blocksize = 4 — ISNR=2.56dB, SSIM=0.693

(d) Blocksize = 8 — ISNR=2.50dB, SSIM=0.690

Figure 9.8: *Restorations with the nonstationary variance model - 40dB BSNR,* $\mathbb{E}^0\left[\sigma_u^2\right] = 10^{-9}$

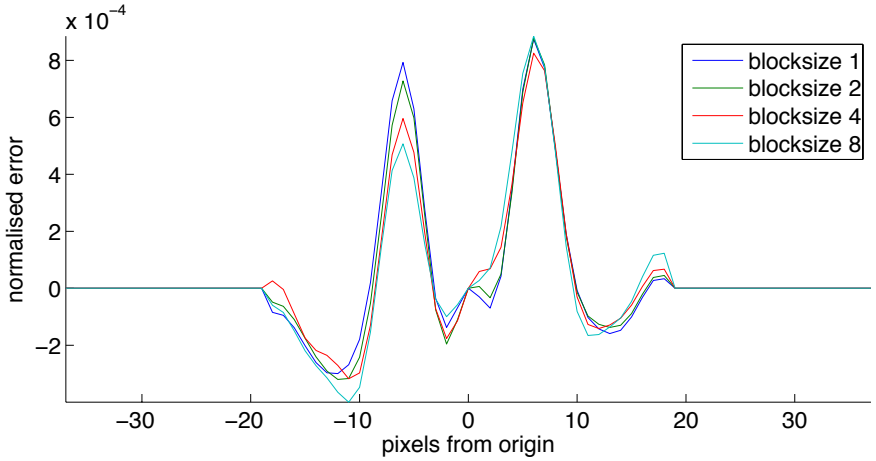


Figure 9.9: *PSF errors, estimated with the nonstationary model — 40dB BSNR,* $\mathbb{E}^0\left[\sigma_u^2\right] = 10^{-8}$

variances which promote stronger application of the prior, and edges have large variances which means they are not smoothed so much by the prior (the effect of noise will be less visible here). This is in a similar spirit to the "classical" spatially adaptive deconvolution techniques, e.g. [21, 110, 121]. However in the experiments we find that there is a limit to the amount of non-stationarity that can be imposed through this method using the VB approach.

We find that decreasing the confidence parameter, $\gamma_{\sigma^2_{v_r}}$, on the variances towards zero, whilst using a reasonable level for the mean of the variances, $\mathbb{E}^0\left[\sigma^2_{v_r}\right]$, (similar to what we estimate from the global variance model) tends to make the estimates become unstable as the iterations progress. That is, the variance of a small number of pixels at corners in the image grows without bound, causing the intensities of these pixels to also become very large in absolute value. This gives the impression of "hot pixels" in the restored image. We mitigate the problem by using larger regions, but this again reduces the amount of local adaptivity we can achieve. We see in fig. 9.8 that the problem goes away with larger block sizes (4 pixel square blocks seems the best compromise).

Unfortunately this problem seems to be a problem of the deterministic nature of the algorithm, and the approximations made in the variational approach, with the distributions we are using. Essentially the inverse Gamma distribution should have a long tail such that very large variances are possible but with small probability. The use of the VB approximation and iterative updates of this mean means that we are not taking into account the moments other than the mean of the distribution. By examining the empirical histograms of the variances estimated from the image at one iteration, as in fig. 9.10, we see that they do not really follow the correct distribution. Two problems are evident: truncation of the left side of the distribution, and bunching of most of the values around this minimum value, $\gamma_{\sigma^2_{v_r}}\mathbb{E}^0\left[\sigma^2_{v_r}\right]$, with fewer than expected pixels extending into the tails.

If we were to instead sample from the distribution to estimate the local variances, these pixels would not suffer from the positive feedback problem as they would be drawn from their true distribution. Moreover, at subsequent iterations, pixels in the upper tail would have a high chance of being sampled at a lower value, thus mitigating the feedback effect. In fig. 9.10(c), we see how the variances from the same approximate distribution ought to look, when sampled.

In the model in the next section, we find that the AR coefficients offer another way of impart-

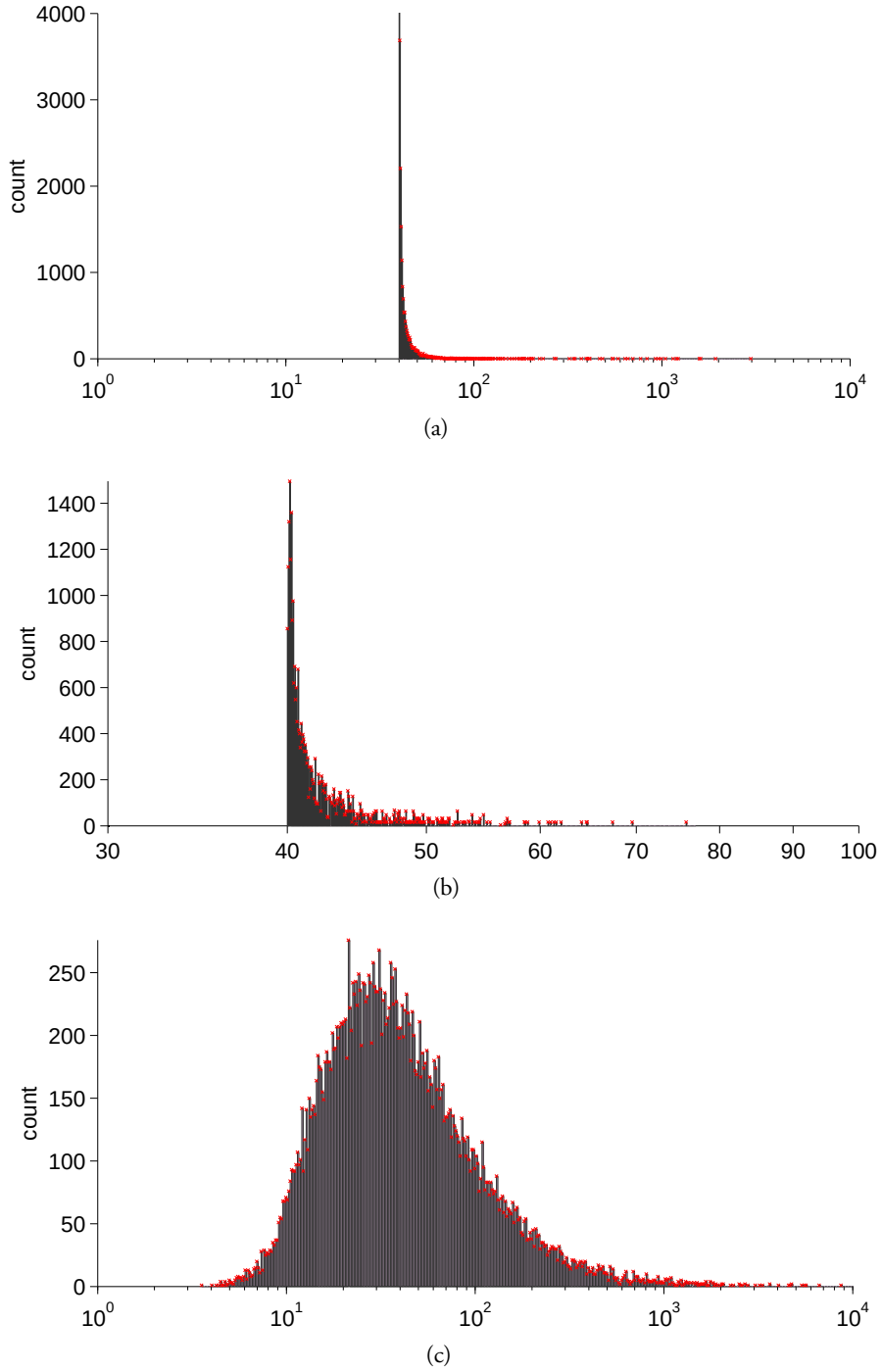


Figure 9.10: *Histograms of image variances* $\{\sigma^2_{\nu_r}\}_{r=1}^{R}$*, found with the VB method and nonstationary variance model: (a) blocksize 1; (b) blocksize 4. (c) Variances sampled from the approximate distribution* $q(\boldsymbol{\sigma}_\nu)$ *corresponding to the case in (a)*

ing a large degree of non-stationarity. Fortunately these coefficients do not suffer to the same extent with the feedback problem.

9.4.3 Use of a nonstationary blur model

One problem with using a stationary model for the blur is that the amount of smoothing for the whole PSF support (represented by the parameter σ_u is determined from an average over the whole blur. Thus increasing the size of the PSF support will tend to incur more smoothing, as the PSF is actually non-homogeneous, with large coefficients at the centre and zero values outside its true support, which for larger support sizes will reduce the term $\frac{1}{L_h}\mathbb{E}\left[\mathbf{u}^T\mathbf{u}\right]$ used in estimating σ_u.

We could consider also using the same nonstationary variance model for the PSF. With a sufficient amount of non-stationarity, the problem just described should be avoided, as the smoothing is estimated locally and thus doesn't change if the support is increased. Ideally this should also help suppress noise in the PSF where it should be zero, by "sparsifying" it. However, in the VB implementation we are not able to specify much nonstationarity as we find a similar problem to the hot pixels in the image occurs:the PSF will tend to a delta function because of positive feedback of the large variances near the origin.

9.5 Nonstationary AR Algorithm

We now arrive at the almost complete model by introducing the AR parameter estimation, which is where we begin to see the real improvements realisable by the new model. This replaces the prior for $\mathbf{f}$ shown in the previous model with the AR prior, as shown in fig. 9.11.

For this model, we find the approximate distributions using the same method as before. In addition to the new distribution $q(\mathbf{f})$, we also need to find the distribution approximation for the AR parameters, $q(\boldsymbol{\alpha}_f)$, and the means $q(\mathbf{m}_\mathbf{f})$. We present the derivations for these expressions in Appendix F. In practice we also use a degenerate distribution for $\boldsymbol{\alpha}_f$, and we find that it is normally equally effective to just use a fixed value of the local means rather than updating at each iteration. That is, we take an estimate of the means of the (blurred) image just at the first iteration, and keep this fixed throughout the algorithm.

Using the degenerate approximation, the update equations for the approximate image distri-

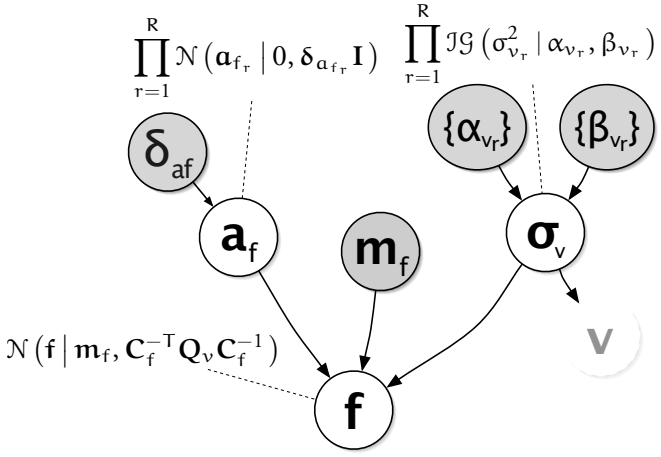


Figure 9.11: *Graphical model for the AR prior*

bution are now found as $q^k(\mathbf{f}) \sim \mathcal{N}(\mathbf{f} \mid \mathbb{E}^k[\mathbf{f}], \text{cov}^k[\mathbf{f}])$ with the parameters

$$\text{cov}^k[\mathbf{f}] = \left(\mathbb{E}^k\left[\sigma_w^{-2}\right] \mathbb{E}^k[\mathbf{H}]^T \mathbb{E}^k[\mathbf{H}] + \|\mathbf{I} - \mathbb{E}^k[\mathbf{A}_f]\|^2_{\mathbf{Q}_v}\right)^{-1} \tag{9.23}$$

$$\begin{aligned}\mathbb{E}^k[\mathbf{f}] = \text{cov}^k[\mathbf{f}] \Bigg(&\mathbb{E}^k\left[\sigma_w^{-2}\right] \mathbb{E}^k[\mathbf{H}]^T \mathbf{g} \\ &+ \left((\mathbf{I} - \mathbb{E}^k[\mathbf{A}_f])^T \mathbb{E}^k\left[\mathbf{Q}_v^{-1}\right] (\mathbf{I} - \mathbb{E}^k[\mathbf{A}_m])\right) \mathbb{E}^k[\mathbf{m}_f] \Bigg).\end{aligned} \tag{9.24}$$

i.e. the same as before except that we use the latest expected value of the AR parameters to define $\mathbf{C}_f = \mathbf{I} - \mathbb{E}^k[\mathbf{A}_f]$ which is used in place of the fixed matrix $\mathbf{C}$. As before we solve for the expected value of the image using CGLS, only this time including the means in the vector on the right hand side (RHS) of the linear system. Thus the system we solve is given by:

$$\mathbf{M}^T \mathbf{M} \mathbf{x} = \mathbf{M}^T \mathbf{y} \tag{9.25}$$

where $\mathbf{Q}_v^{-1} = \mathbf{L}^T \mathbf{L}$, $\mathbf{M} = \begin{bmatrix} \frac{\sigma_w^{-1} \mathbf{H}}{\mathbf{L} \mathbf{C}_f} \end{bmatrix}$, and $\mathbf{y} = \begin{bmatrix} \frac{\sigma_w^{-1} \mathbf{g}}{\mathbf{L} \boldsymbol{\eta}_f} \end{bmatrix} + \boldsymbol{\nu}$.

The approximate distribution for $q(\mathbf{a}_f)$ is then found independently in each region as: $q^k(\mathbf{a}_{f_r}) \sim \mathcal{N}(\mathbf{a}_{f_r} \mid \mathbb{E}^k[\mathbf{a}_{f_r}], \text{cov}^k[\mathbf{a}_{f_r}])$, with

$$\text{cov}^k[\mathbf{a}_{f_r}] = \left(\|\mathbb{E}^k[\mathbf{F}_{a_r}] - \mathbb{E}^k[\mathbf{M}_{a_r}]\|^2 + \delta_a^{-1} \mathbf{I}_{L_a}\right)^{-1} \tag{9.26}$$

$$\mathbb{E}^k[\mathbf{a}_{f_r}] = \text{cov}^k[\mathbf{a}_{f_r}] \left((\mathbb{E}^k[\mathbf{F}_{a_r}] - \mathbb{E}^k[\mathbf{M}_{a_r}])^T (\mathbb{E}^k[\mathbf{f}_r] - \mathbb{E}^k[\mathbf{m}_{f_r}])\right). \tag{9.27}$$

The expected value of the AR parameters in each block can be solved directly (by Gaussian elimination) as each system is relatively small; however as this has to be repeated for each block in the image the speed is important, and we again find it is slightly faster to use a few iterations of conjugate gradients to solve (9.27), in a similar manner to how we find the expected value of the image.

9.5.1 Examples and effect of choice of segmentation

In fig. 9.12 and fig. 9.13 we present restorations using the nonstationary AR model of the same degraded images as in the previous section, at 40dB and 20dB BSNR. In each row we present on the right the restoration at the iteration that gave the best SSIM value, and on the left that at the first iteration where the SSIM exceeded 98% of this value. What this shows is that a visually acceptable restoration is achieved much sooner than these figures would suggest. As the iterations progress, actually the visual appearance tends to degrade, even if the quality metrics are increasing, mainly due to lack of contrast in the image (this is the main difference observable in the images on the right; note that the levels in the image are adjusted to show the full dynamic range, thus any small local highlights or shadows due to ringing will reduce the overall contrast). This effect is more evident when viewing all the iterations in the movie clips of the iterations (see Appendix H).

In each figure we also show in each row how the choice of the regions at each iteration affects the results. In the first row of each figure, we show the results using the model as presented: using a fixed grid of blocks. We have chosen 8×8 pixel blocks, based on previous evidence that they can model the signal well, and as the best trade-off between quality and complexity (as the size decreases, more AR parameters must be estimated and stored, increasing run time and memory usage).

However with this choice, although the results are satisfactorily sharp, artefacts at block boundaries in the form of discontinuities can appear (note the size of these artefacts depend also somewhat on the choice of the parameter δ_{α_f}, see §9.6). We could consider a model that is a uniform mixture of different AR models defined on overlapping grids. Updating all the models at each iteration would drastically increase computations. Instead, we propose modifying the algorithm, such that at each iteration, one of the grids (i.e. a random offset of the starting grid) is chosen. The results using all 64 possible 8×8 grids are shown in the second rows of the figures. Note that the block artefacts are gone, but also the image has lost some sharpness.

(a) Fixed blocks, After 548 Iterations — ISNR=2.40dB, SSIM=0.691

(b) Fixed blocks, After 2486 Iterations — ISNR=3.00dB, SSIM=0.705

(c) Fully overlapping blocks, update every iteration; After 587 Iterations — ISNR=2.22dB, SSIM=0.683

(d) Fully overlapping blocks, update every iteration; After 2072 Iterations — ISNR=2.76dB, SSIM=0.697

(e) Half overlapping blocks, update every 20 iterations, After 380 Iterations — ISNR=3.00dB, SSIM=0.707

(f) Half overlapping blocks, update every 20 iterations; After 940 Iterations — ISNR=3.01dB, SSIM=0.713

Figure 9.12: *Restorations with the nonstationary AR model - 40dB BSNR*

(a) Fixed blocks, After 14 Iterations — ISNR=1.53dB, SSIM=0.614

(b) Fixed blocks, After 274 Iterations — ISNR=2.07dB, SSIM=0.625

(c) Fully overlapping blocks, update every iteration; After 26 Iterations — ISNR=1.79dB, SSIM=0.630

(d) Fully overlapping blocks, update every iteration; After 132 Iterations — ISNR=2.26dB, SSIM=0.642

(e) Half overlapping blocks, update every 20 iterations, After 141 Iterations — ISNR=1.85dB, SSIM=0.633

(f) Half overlapping blocks, update every 20 iterations; After 660 Iterations — ISNR=2.32dB, SSIM=0.645

Figure 9.13: *Restorations with the nonstationary AR model - 20dB BSNR*

In the third rows, we find what appears the best option, to only use four different grids, with half a block overlap, and to only pick a new grid at random every 20 iterations. This last step helps maintain some sharpness as it usually requires several iterations of the image in the same grid location before the AR parameters begin to converge. Use of only four grids rather than 64 also saves memory, as in the implementation we pre-cache the structure of the resulting matrices $\mathbf{C}$ based on the grids, in order to calculate them more rapidly.

Another simple strategy that was considered was using a quad-tree segmentation to provide the regions, based upon homogeneity of the regions' texture. However this method was not pursued further, as the overlapping blocks approach is simple and seems effective enough. A full arbitrary segmentation of the image to provide the regions would also be possible, but it may be at much additional cost, for a small gain in performance.

9.6 Testing the new prior: non-blind restoration

To see how effective the new prior can be we also show some results using the same restoration procedure when the PSF is known. This is also useful to investigate the effect of the choice of the hyper-hyperparameter δ_{α_f}. With the non-blind restorations the scale and shift ambiguities are not a problem, as we have the original PSF, so it is more valid to compare ISNR and SSIM figures for different settings.

We run the same VB algorithm using the nonstationary AR prior, with the same settings as before only omitting the blur estimation stage (other hyperparameters are estimated as usual, and we use half overlapping blocks with updates of the grid every 20 iterations). Four experiments are considered: the two cases shown already with a 3 pixel standard deviation Gaussian PSF at 20 and 40dB BSNR, and a 12 pixel diameter circular PSF, also with noise at 20 and 40dB. The deconvolution results in each case, for various different choices of δ_{α_f} are shown in fig. 9.14–9.17.

In fig. 9.18(a) and 9.18(b), we show the ISNR and SSIM results for these experiments. The horizontal lines show a reference level from the stationary restorations (using the same VB method to estimate the unknown hyperparameters). We can see that a peak in all cases appears at around or just above $\delta_{\alpha_f} = 3 \times 10^{-3}$. For this setting, the nonstationary model significantly outperforms the stationary model in all cases; hence we have chosen to use this fixed value for all the blind restorations in this chapter. Also note the SSIM graphs indicate we will loose

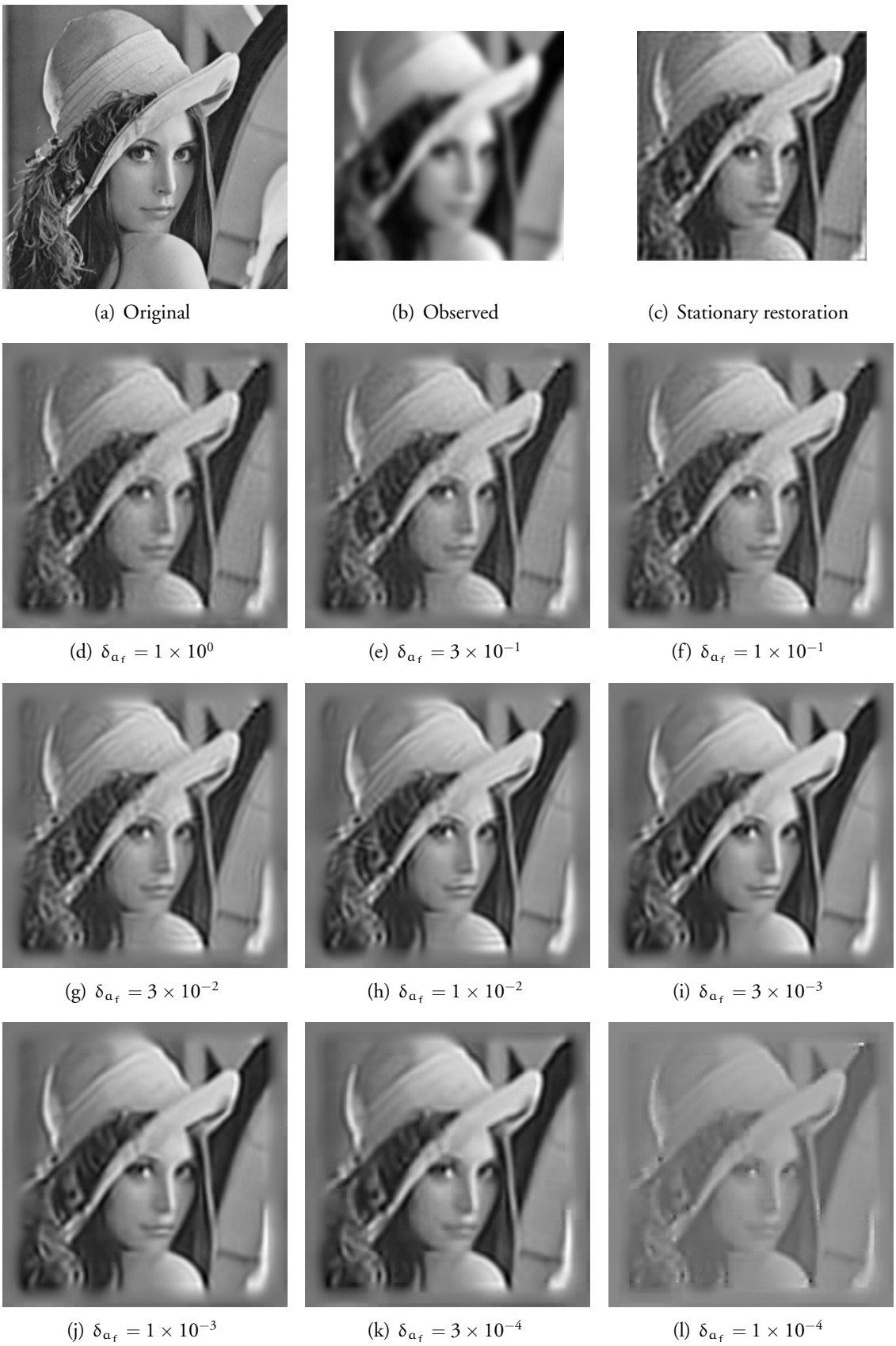
more by overestimating this parameter rather than underestimating. When we use the model with the MCMC method in the next chapter, we find that it is sometimes best to use a slightly smaller value, e.g. $\delta_{\alpha_f} = 1 \times 10^{-3}$. This is because there is a chance that the filters can become unstable if less regularised, in a similar manner to the way the variances did in the last section with the "hot pixel" problem. This is more of a problem when we introduce extra noise into the system as the Gibbs sampler does, and when we use higher temperatures.

Considering the graphs in fig. 9.18(a) and 9.18(b) again, the nonstationary model also shows a bigger improvement over the stationary model in the circular blur case; this is because it is possible to recover more high frequencies than in the Gaussian blur case when they are completely lost. The 20dB results show bigger disagreement between the two metrics: ISNR suggests using a larger δ_{α_f} isn't so bad as SSIM reports — this is where there is extreme noise amplification by over-fitting the model. However there is still an agreement around the optimum values.

Referring to the visual results helps to illustrate more clearly what is going on. We make a number of observations. Especially in the 40dB cases, as δ_{α_f} grows to 1, the restorations become very similar to the stationary restoration, where there is much ringing due to partial regularisation of the noise and partial deblurring. These images are slightly sharper than the stationary case in the iterations shown, but get even more sharper and noisier as the iterations progress (again we have shown the images at the iteration giving peak SSIM — see the videos in Appendix H to observe the full evolution).

As we decrease δ_{α_f}, the noise starts to get suppressed, until at around a value of 3×10^{-2} it resembles ringing just around the edges. For $\delta_{\alpha_f} = 3 \times 10^{-3}$, the ringing disappears although we still have a sharp image. As it decreases further, some smaller details in the image are lost and the image becomes progressively smoother, until at around $\delta_{\alpha_f} = 1 \times 10^{-4}$ the "hot pixel" effect returns.

This is explained nicely by considering the magnitude of the coefficients. We have two variables, $\boldsymbol{\alpha}_{f_r}$ and σ_{v_r}, both in a sense responsible for representing the local activity in each region. For large values of δ_{α_f}, we are not regularising $\boldsymbol{\alpha}_{f_r}$ and therefore the AR filters will adapt freely to whatever signal they find, including the noise (or an approximation of it). In this case we observe that the histograms of the variances σ_{v_r} are clustered around a single value, i.e. they are not used to model the nonstationarity at all. In the opposite case, for very small δ_{α_f}, the AR coefficients are heavily regularised, with many of them taking values close to zero. This makes



(a) Original (b) Observed (c) Stationary restoration

(d) $\delta_{\alpha_f} = 1 \times 10^{0}$ (e) $\delta_{\alpha_f} = 3 \times 10^{-1}$ (f) $\delta_{\alpha_f} = 1 \times 10^{-1}$

(g) $\delta_{\alpha_f} = 3 \times 10^{-2}$ (h) $\delta_{\alpha_f} = 1 \times 10^{-2}$ (i) $\delta_{\alpha_f} = 3 \times 10^{-3}$

(j) $\delta_{\alpha_f} = 1 \times 10^{-3}$ (k) $\delta_{\alpha_f} = 3 \times 10^{-4}$ (l) $\delta_{\alpha_f} = 1 \times 10^{-4}$

Figure 9.14: *Effect of varying δ_{α_f} for the 40dB BSNR Gaussian blurred image*

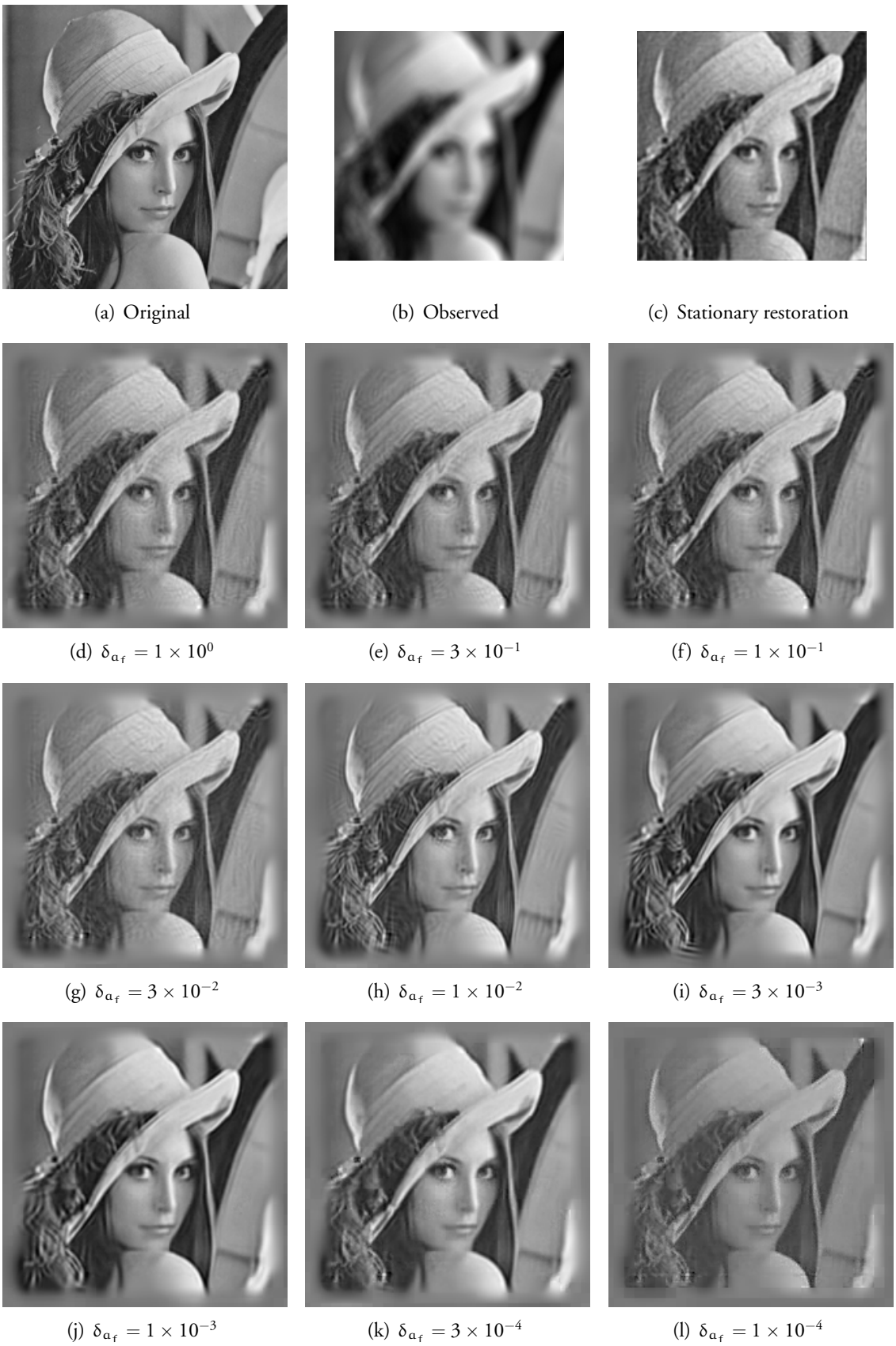
(a) Original (b) Observed (c) Stationary restoration

(d) $\delta_{\alpha_f} = 1 \times 10^{0}$ (e) $\delta_{\alpha_f} = 3 \times 10^{-1}$ (f) $\delta_{\alpha_f} = 1 \times 10^{-1}$

(g) $\delta_{\alpha_f} = 3 \times 10^{-2}$ (h) $\delta_{\alpha_f} = 1 \times 10^{-2}$ (i) $\delta_{\alpha_f} = 3 \times 10^{-3}$

(j) $\delta_{\alpha_f} = 1 \times 10^{-3}$ (k) $\delta_{\alpha_f} = 3 \times 10^{-4}$ (l) $\delta_{\alpha_f} = 1 \times 10^{-4}$

Figure 9.15: *Effect of varying δ_{α_f} for the 20dB BSNR Gaussian blurred image*



(a) Original (b) Observed (c) Stationary restoration

(d) $\delta_{\alpha_f} = 1 \times 10^{0}$ (e) $\delta_{\alpha_f} = 3 \times 10^{-1}$ (f) $\delta_{\alpha_f} = 1 \times 10^{-1}$

(g) $\delta_{\alpha_f} = 3 \times 10^{-2}$ (h) $\delta_{\alpha_f} = 1 \times 10^{-2}$ (i) $\delta_{\alpha_f} = 3 \times 10^{-3}$

(j) $\delta_{\alpha_f} = 1 \times 10^{-3}$ (k) $\delta_{\alpha_f} = 3 \times 10^{-4}$ (l) $\delta_{\alpha_f} = 1 \times 10^{-4}$

Figure 9.16: *Effect of varying δ_{α_f} for the 40dB BSNR circular-shaped PSF blurred image*

(a) Original (b) Observed (c) Stationary restoration

(d) $\delta_{\alpha_f} = 1 \times 10^{0}$ (e) $\delta_{\alpha_f} = 3 \times 10^{-1}$ (f) $\delta_{\alpha_f} = 1 \times 10^{-1}$

(g) $\delta_{\alpha_f} = 3 \times 10^{-2}$ (h) $\delta_{\alpha_f} = 1 \times 10^{-2}$ (i) $\delta_{\alpha_f} = 3 \times 10^{-3}$

(j) $\delta_{\alpha_f} = 1 \times 10^{-3}$ (k) $\delta_{\alpha_f} = 3 \times 10^{-4}$ (l) $\delta_{\alpha_f} = 1 \times 10^{-4}$

Figure 9.17: *Effect of varying δ_{α_f} for the 20dB BSNR circular-shaped PSF blurred image*

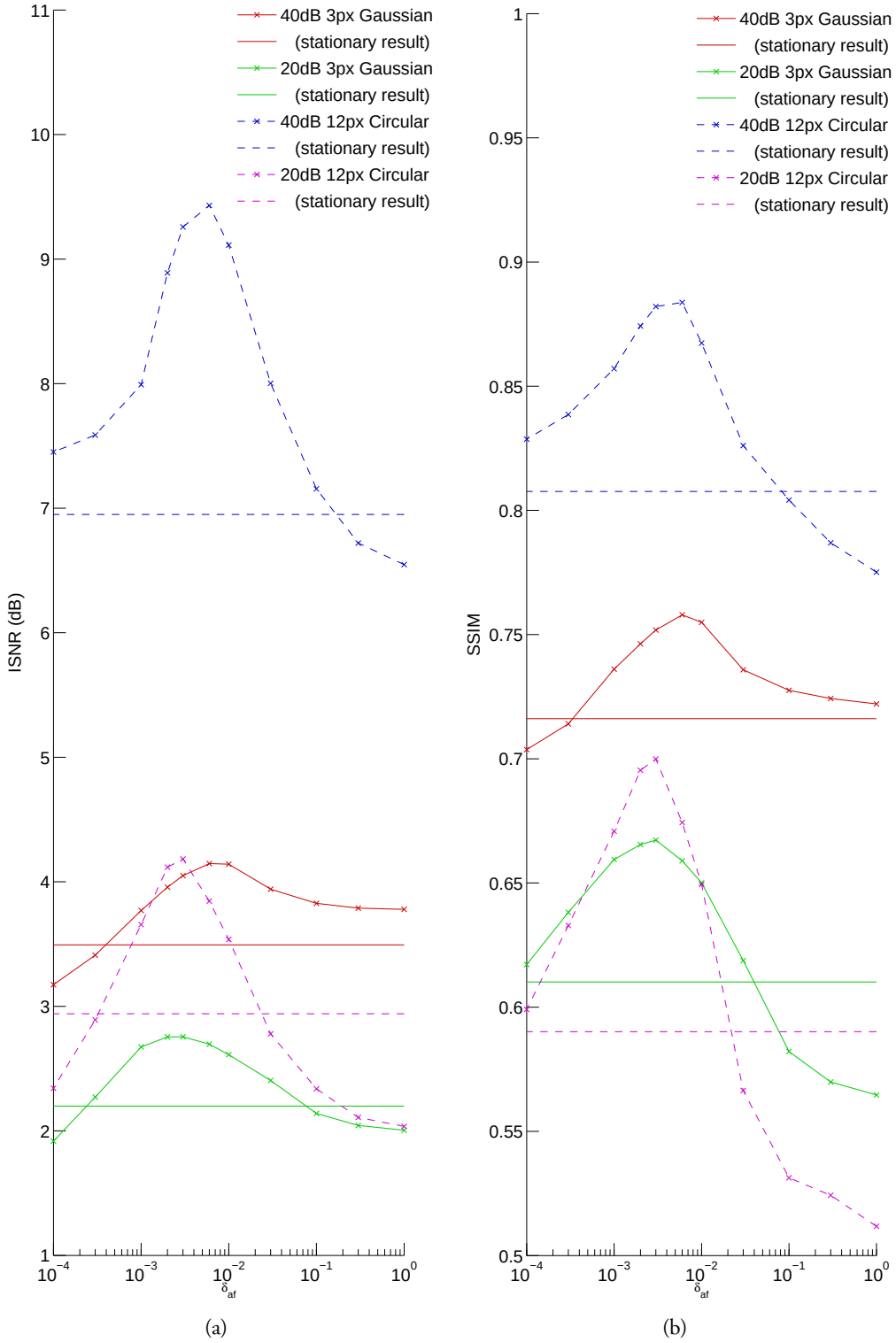


Figure 9.18: *(a) ISNR and (b) SSIM versus various values of δ_{af}, shown for 4 different experiments, as described in the main text. Horizontal lines indicate results using the stationary prior.*

the filters in $\mathbf{C}_f$ almost equal to the identity, and we are only regularising the magnitude of the image, but allowing strong edges to still be preserved. However, in this case, the variances σ_{v_r} become more important in modelling the local activity, and we see their histograms are more spread out. This is the reason why we see the "hot pixel" effect described in the nonstationary variance model again in this case.

Finally we note that even though the numerical results are encouraging, and we achieve very good restorations when there is sufficient information in the blurred image, some other cases such as the 20dB BSNR circular blur case are still challenging for any restoration method, and it is not clear that any of the results in this case are extracting all the information in the blurred image to give a visually pleasing result. The $\delta_{\alpha_f} = 1 \times 10^{-2}$ case in fig. 9.17(h) appears to have the most detail, but there are quite a lot of spurious textures appearing in smooth regions. One suggestion that needs further investigation is that we could add another stage to the hierarchical model, with a prior for δ_{α_f}, and estimate its values locally. The mean and variance of likely values of this prior could be estimated from these experiments, and hopefully we would see that in smooth regions of the image δ_{α_f} would take on low values to promote smoothing, and in regions with more structure, higher values would enable recovery of the texture, thus giving a double-layered spatial adaptivity.

9.7 Conclusions

In this chapter, the variational Bayesian method of posterior approximation has been discussed, and applied to the BID problem. We reviewed how it was used in [153], before developing the stationary models into nonstationary ones. Firstly we considered a nonstationary variance model for the image, which although nice in theory seems to have some practical limitations when used with the iterative VB method.

Subsequently, the nonstationary AR image model was investigated using VB to estimate the required parameters. In this work we have used deterministic approximations of the distributions, however with the AR model we find very good results are obtainable with some substantial improvements over the stationary case. We showed blind restorations in the case of Gaussian blur, and also considered the performance of model under different settings in the non-blind case. This enables us to consider good values for the hyper-hyperparameter δ_{α_f}, which controls how closely the model fits the restoration, and thus the amount of adaptive

noise removal is done.

In terms of the results, we have demonstrated both improvements of the new AR image model in non-blind image restorations (see the figures in §9.6), but we have also shown the prior leads to improved blind restorations, over the method of [153] for the Gaussian blur (compare for instance fig. 9.3(d) with fig. 9.12(f)). For other types of blur, the VB method presented here may not converge well for blind restoration; however further gains are shown using the new prior with the MCMC method in the next chapter.

The AR model can restore not only a piecewise smooth signal, but also sharp edges, and most importantly regions of texture. We are not aware of other models that have been proposed for blind deconvolution that can restore texture in this spatially adaptive manner, rather than just preserving edges and smooth regions. However, there are many other more advanced statistical image models that have been proposed for simpler problems such as image denoising. These include multiscale, transform domain methods, e.g. [177], and methods based on learned dictionaries, e.g. [64]. It would be interesting to try to incorporate some of these features to the proposed model. For instance we could consider learning distributions of the AR parameters, as well as using a multi-scale version. As an alternative model, we could also consider ways to solve the problem with the instability of the variance estimation when the AR model is not present, by modelling correlations between the variances themselves.

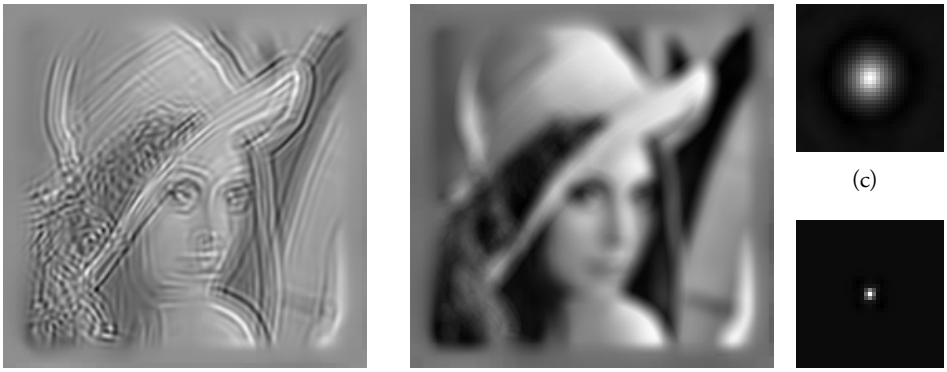
Chapter 10
Gibbs Sampler

We have shown that the models presented so far can provide a useful solution to the deconvolution, and in certain cases, blind image deconvolution problems. For instance, the smoothness prior for the PSF is appropriate in the case of Gaussian blur, and we have shown good results in this case. However it is important to remember that the VB method, even without the degenerate approximations we have used, is only locally convergent, and depending on both the initial conditions and the accuracy of the priors, may not even converge at all.

This is what we observe if we try to use the previous algorithm when the blur used is, for example an out-of-focus blur (circular PSF). The stationary smoothness prior for the PSF does not well model the sharp edges in the PSF at all. The deterministic nature of the iterative algorithm tends to mean that it diverges from the start, because the ringing artefacts in the image due to error in the PSF become re-enforced as false edges in the image (see fig. 10.1).

One possibility is to use the same nonstationary AR model we used for the image prior. The nonstationary AR model can well model the sharp edges of the circular PSF, however there are other stationary points of the IVB algorithm which will be met first if we initialise with a Gaussian as before. Experiments using this model also failed to converge to the correct solution, instead converging to approximately a delta function for the PSF (see fig. 10.1). We could try to find a better initialisation, although then the method looses generality. In our experiments we were unable to find an initialisation or setting of the known parameters that led to convergence. Use of a different model for the PSF, for example a piecewise constant model such as TV [9, 50, 236], can help convergence for these types of blurs. However this does not apply well to other types of PSF which are not piecewise constant, so unfortunately generalisation is also lost unless some kind of model selection is performed to choose the best PSF model. We may also want to consider many other types of PSF such as non-constant motion blur or lens aberrations, which are not very well constrained by any model yet proposed.

Using stochastic methods that do a more thorough exploration of the parameter space is one way to get around these problems. We have given a brief account of how sampling based



(a) (b) (c) (d)

Figure 10.1: *Unsuccessful blind restorations of the 12 pixel diameter circular blurred image, using the VB approach and (a) Stationary and (b) nonstationary AR prior models for the PSF; (c) and (d) their estimated PSFs.*

Markov chain Monte Carlo (MCMC) methods may be applied to the blind deconvolution problem in §5.2.6. We choose to adopt this approach in this chapter, and show how we get the method to work in practice.

MCMC methods have been used successfully in many 1D signal processing problems, including blind deconvolution for audio and communications, as well as image processing problems in 2D such as segmentation and non-blind deconvolution [76]. However one reason that it seems they have not previously been applied to BID is because of the dimension of the problem: as we have previously explained, it is computationally intractable to invert the covariance matrix of the image PDF due to its dimension. 1D methods typically require factorisation of this covariance in order to draw samples efficiently. Previous MCMC image restoration approaches, e.g. [76], sampled the image pixelwise. However this results in slow convergence, especially when we have to sample the whole image many times. In our method we use bulk sampling of the image, along with simulated annealing (SA) to explore the solution space more rapidly and speed up convergence, making the solution become possible with a moderate amount of computational effort.

In the next section, we discuss some details of how we can use the Gibbs sampler for BID. Then in the experiments in this chapter, we show results using the proposed MCMC method for a number of different blurs. We show that the method can converge satisfactorily under several different conditions, and that we can obtain state of the art results.

10.1 Overview of MCMC approach

Recall that we have formulated the problem to be solved in a general hierarchical Bayesian framework. The goal is to use the posterior distribution of the unknowns $\boldsymbol{\Theta}$ conditional on the observed data, $p(\boldsymbol{\Theta} \mid \mathbf{g}) \propto p(\mathbf{g} \mid \boldsymbol{\Theta})\, p(\boldsymbol{\Theta})$ to perform inference about the likely values of $\boldsymbol{\Theta}$, or some subset of it. For all the parameters in our complete model presented in §6.7, we have $\boldsymbol{\Theta} = \{\mathbf{f}, \boldsymbol{\alpha}_f, \boldsymbol{\sigma}_v, \boldsymbol{\mu}_f, \mathbf{h}, \boldsymbol{\alpha}_h, \boldsymbol{\sigma}_u, \boldsymbol{\mu}_h, \sigma_w^2\}$.

With such a large parameter space for $\boldsymbol{\Theta}$, it is infeasible to perform direct optimisation or calculation of this function. We have shown one way to reduce the dimension of the space is to marginalise nuisance variables we are not interested in, however with the full model this also becomes intractable. As we have just described, approximation methods may fail to converge to the correct solution. Therefore we would like a method that can solve the problem of parameter estimation in high dimensional spaces. It turns out that MCMC methods (see §5.2.6) are particularly suitable for these types of problems.

As discussed in §5.2.6, applying the Gibbs sampler to a problem formulated in the hierarchical Bayesian framework is particularly straightforward, so long as the required conditional distributions can be found in a standard form. As it turns out, with our tree-structured models, these conditionals are the exact same distributions that we use as the approximate distributions in the VB approach.

Therefore transitioning to using the Gibbs sampler requires only that we can sample these conditionals rather than evaluating their expectations. This makes design of the new algorithms simple; the only new task we need to do is find out how to sample from these conditional distributions. Fortunately because we have used conjugate priors, all of the conditionals are already in the nice form of Gaussian or inverse-Gamma distributions. However there are a couple of tricks we need to use in order to make the process work effectively.

10.2 Applying the Gibbs Sampler to the BID problem

In this section we consider using the Gibbs sampler to sample the full posterior distribution defined in §6.7. This is as we used before at the end of the last chapter, except we now also use the AR model for the PSF, and we introduce positivity constraints on the PSF and means. It was found experimentally that adding in these positivity constraints ensures much better

convergence towards the true PSF, and helps remove a lot of the ringing artefacts in the image, however the manner in which to do so had to be investigated carefully.[1] We propose modifying the Gaussian distributions to rectified Gaussians to impose this constraint.

Although the basic method of application of the Gibbs sampler, described in §5.2.6.3 is particularly straightforward, and appears to be just a slight modification of the previous VB algorithms, we need to consider carefully how to draw the samples. Several issues that had to be overcome in practice to make the Gibbs sampler work for the BID problem. These will be discussed in this section, and include:

- Bulk-sampling of the image
- Positivity of the PSF
- Improving convergence speed.

The Gibbs sampler draws samples $\theta^{(k)}$ at iteration k from the distribution of each $\theta \in \boldsymbol{\Theta}$ in turn, conditional on the previously sampled values of all the other unknowns. These samples form a Markov chain, whose stationary distribution is the desired posterior.

These conditional PDFs, $p\left(\theta \,|\, \boldsymbol{\Theta}_{\setminus\theta}\right)$, only need be known up to a constant of proportionality; hence they are calculated simply by taking the product of the terms from $p\left(\mathbf{g} \,|\, \boldsymbol{\Theta}\right) p\left(\boldsymbol{\Theta}\right)$ that contain the variable θ of interest. In the tree-structured hierarchical model, this corresponds to the product of the prior and the child variable's distribution. Applying this procedure to the full model in fig. 6.8 results in the conditionals shown in Algorithm 2, where we also define the sets of image and blur parameters $\boldsymbol{\theta}_{f_r} = \{L_{r_+}, \alpha_{v_r}, \beta_{v_r}, \delta_{\mu_f}, \delta_{a_{f_r}}, L_{a_f}\}$ and $\boldsymbol{\theta}_{h_r} = \{L_{h_{r_+}}, \alpha_{u_r}, \beta_{u_r}, \delta_{\mu_h}, \delta_{a_{h_r}}, L_{a_h}\}$.

10.2.1 Sampling Procedure for Image and Blur

While the scalar Gaussian and IG distributions are easily sampled from standard pseudo-random number generators (the IG by the reciprocal of Gamma distributed samples), the multivariate Gaussians for the image and blur should be treated more carefully.

[1] This positivity could be applied by a constrained minimisation method or POCS after sampling, however in our experience this tends to actually make the algorithm become unstable, as we are not sampling from the correct distribution.

Input: $\mathbf{g}, \mathbf{h}^{(0)}, \theta_{f_r}, \theta_{h_r}$
Output: $\hat{\mathbf{f}} = \frac{1}{K-k_0}\sum_{k=k_0}^{K}\mathbf{f}^{(k)}, \hat{\mathbf{h}} = \frac{1}{K-k_0}\sum_{k=k_0}^{K}\mathbf{h}^{(k)}$

for $k = 1 \cdots n$ **do**
  offset $\mathcal{R}_r, \forall r = 1 \cdots R$, by $(i,j) \leftarrow \text{Uniform}(m_r)$;
  $\boldsymbol{w}^{(k)} = \mathbf{g} - \mathbf{H}\mathbf{f}$;
  $\left(\sigma_w^2\right)^{(k)} \leftarrow \mathcal{IG}\left(\sigma_w^2 \,\middle|\, \alpha_w + \frac{L_+}{2}, \beta_w + \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}\right)$;
  for $r = 1 \cdots R$: **do**
    $\{\left(\sigma_{v_r}^2\right)^{(k)}, \mu_{f_r}^{(k)}, \mathbf{a}_{f_r}^{(k)}\} =$`ParamSampling`$(r, \sigma_w^2, \mathbf{C}_{f_r}, \mathbf{f}_{r_+}, \mu_{f_r}, \theta_{f_r})$;
  $\left(\text{cov}^k[\mathbf{f}]\right)^{-1} = \mathbf{C}_f^T\mathbf{Q}_v^{-1}\mathbf{C}_f + \sigma_w^{-2}\mathbf{H}^T\mathbf{H}$;
  $\mathbb{E}^k[\mathbf{f}] = \text{cov}^k[\mathbf{f}]\left(\sigma_w^{-2}\mathbf{H}^T\mathbf{g} + \mathbf{C}_f^T\mathbf{Q}_v^{-1}\boldsymbol{\eta}_f\right)$;
  $\mathbf{f}^{(k)} \leftarrow \mathcal{N}(\mathbf{f} \mid \mathbb{E}^k[\mathbf{f}], \text{cov}^k[\mathbf{f}])$;
  offset $\mathcal{R}_{h_r}, \forall r = 1 \cdots R_h$, by $(i,j) \leftarrow \text{Uniform}(m_{r_h})$;
  for $r = 1 \cdots R_h$: **do**
    $\{\left(\sigma_{u_r}^2\right)^{(k)}, \mu_{h_r}^{(k)}, \mathbf{a}_{h_r}^{(k)}\} =$`ParamSampling`$(r, \sigma_w^2, \mathbf{C}_{h_r}, \mathbf{h}_{r_+}, \mu_{h_r}, \theta_{h_r})$;
  $\left(\text{cov}^k[\mathbf{h}]\right)^{-1} = \mathbf{C}_h^T\mathbf{Q}_v^{-1}\mathbf{C}_h + \sigma_w^{-2}\mathbf{F}^T\mathbf{F}$;
  $\mathbb{E}^k[\mathbf{h}] = \text{cov}^k[\mathbf{h}]\left(\sigma_w^{-2}\mathbf{F}^T\mathbf{g} + \mathbf{C}_h^T\mathbf{Q}_u^{-1}\boldsymbol{\eta}_h\right)$;
  $\mathbf{h}^{(k)} \leftarrow \mathcal{N}^+(\mathbf{h} \mid \mathbb{E}^k[\mathbf{h}], \text{cov}^k[\mathbf{h}])$;

Algorithm 2: *Gibbs sampling for Blind Image Deconvolution*.
The symbol $\leftarrow$ means draw a sample from the RHS distribution. The indices k or $k-1$ of the quantities in the RHS of the equations have been dropped for clarity; it is implied that the most recently sampled value is used in each case.

We could take the approach of previous methods in image processing such as [76], and sample the image pixel by pixel, based upon the conditional distributions of each pixel in the MRF. However this not only results in in more work due to the evaluation of extra `for` loops in the code, but also makes the Markov chain *mix*, or traverse the parameter space more slowly. This happens whenever there is strong correlation between the variables. As a simple example, consider sampling a 2D Gaussian with a strong correlation coordinate-wise. Rather than moving freely about the distribution, each step will be trapped in one part of the density, as we only move along one coordinate axis at a time in the region of high probability, requiring a stair-step type movement in order to traverse the major axis of the Gaussian. Clearly it is better if we know these variables are correlated that we can sample them jointly.

Typically a multivariate Gaussian is sampled jointly by finding a Cholesky decomposition of the covariance matrix, and multiplying this by a white Gaussian $\mathcal{N}(\cdot \mid \mathbf{0}, \mathbf{I})$ of appropriate dimension, followed by adding on the mean of the conditional. However this is intractable for such a large matrix as $\text{cov}^k[\mathbf{f}]$. Instead, a CGLS solver is used as before, but with a modification to include the random component [101]. For example, $\mathbf{f}$ is sampled avoiding calculation of

Input: $r, \sigma_w^2, \mathbf{C}_{f_{r_+}}, \mathbf{f}_{r_+}, \mu_{f_r}, \theta_{f_r}$
Output: $\{(\sigma_{\nu_r}^2)^{(k)}, \mu_{f_r}^{(k)}, \mathbf{a}_{f_r}^{(k)}\}$

$\boldsymbol{\nu}_r^{(k)} = \mathbf{C}_{f_{r_+}}(\mathbf{f}_{r_+} - \mathbf{m}_{f_{r_+}})$;
$(\sigma_{\nu_r}^2)^{(k)} \leftarrow \mathcal{IG}\left(\sigma_{\nu_r}^2 \,\middle|\, \alpha_{\nu_r} + \frac{L_r}{2}, \beta_{\nu_r} + \frac{1}{2}\boldsymbol{\nu}_r^T\boldsymbol{\nu}_r\right)$;
$\text{cov}^k[\mu_{f_r}] = \left(\|\mathbf{C}_{f_{r_+}}\mathbf{1}_{L_{r_+}}\|^2/\sigma_{\nu_r}^2 + \delta_{\mu_f}^{-1}\right)^{-1}$;
$\mathbb{E}^k[\mu_{f_r}] = \frac{\mathbf{1}_{L_{r_+}}^T\mathbf{C}_{f_{r_+}}^T\mathbf{C}_{f_{r_+}}\mathbf{f}_{r_+}/\sigma_{\nu_r}^2}{\|\mathbf{C}_{f_{r_+}}\mathbf{1}_{L_{r_+}}\|^2/\sigma_{\nu_r}^2 + \delta_{\mu_f}^{-1}}$;
$\mu_{f_r}^{(k)} \leftarrow \mathcal{N}^+(\mu_{f_r} \mid \mathbb{E}^k[\mu_{f_r}], \text{cov}^k[\mu_{f_r}])$;
$\text{cov}^k[\mathbf{a}_{f_r}] = \left(\mathbf{F}_{a_r}^{*\,T}\mathbf{F}_{a_r}^*/\sigma_{\nu_r}^2 + \delta_{a_{f_r}}^{-1}\mathbf{I}_{L_{a_f}}\right)^{-1}$;
$\mathbb{E}^k[\mathbf{a}_{f_r}] = \text{cov}^k[\mathbf{a}_{f_r}]\,\sigma_{\nu_r}^{-2}\mathbf{F}_{a_r}^{*\,T}(\mathbf{f}_r - \mathbf{m}_{f_r})$;
$\mathbf{a}_{f_r}^{(k)} \leftarrow \mathcal{N}(\mathbf{a}_{f_r} \mid \mathbb{E}^k[\mathbf{a}_{f_r}], \text{cov}^k[\mathbf{a}_{f_r}])$;

Function `ParamSampling`

$\text{cov}^k[\mathbf{f}]$ and $\mathbb{E}^k[\mathbf{f}]$ by drawing $\boldsymbol{\nu} \leftarrow \mathcal{N}\left(\boldsymbol{\nu} \,\middle|\, \mathbf{0}_{L+L_+}, \mathbf{I}_{L+L_+}\right)$ and iteratively solving

$$\mathbf{M}^T\mathbf{M}\boldsymbol{x} = \mathbf{M}^T\boldsymbol{y} \tag{10.1}$$

where $\mathbf{Q}_v^{-1} = \mathbf{L}^T\mathbf{L}$, $\mathbf{M} = \begin{bmatrix} \sigma_w^{-1}\mathbf{H} \\ \mathbf{L}\mathbf{C}_f \end{bmatrix}$, and $\boldsymbol{y} = \begin{bmatrix} \sigma_w^{-1}\mathbf{g} \\ \mathbf{L}\boldsymbol{\eta}_f \end{bmatrix} + \boldsymbol{\nu}$. Note this is exactly the same as in the VB approach, apart from the addition of $\boldsymbol{\nu}$.

The same approach is used to sample $\mathbf{a}_{f_r}$ and $\mathbf{a}_{h_r}$, though we make the approximation in $p(\mathbf{f} \mid \mathbf{a}_f, \sigma_v, \mathbf{m}_f)$ of ignoring the dependency of the normalising term $\det|\boldsymbol{\Sigma}_{f_+}^{-1}|$ on $\mathbf{a}_f$, resulting in an improper prior. This does not seem to adversely affect results however.

10.2.2 Sampling rectified Gaussians

To sample the rectified Gaussians, rejection sampling is used fig. 10.2. The principle is to use a proposal distribution $q(\cdot)$ that when multiplied by some constant, $M < \infty$, upper bounds the true distribution $q(\cdot)$ for all $x \in \mathcal{X}$. We need to be able to sample $\{x^{(r)}\}_{r=1}^R \leftarrow q(x)$, to evaluate $q(\cdot)$ and $q(\cdot)$, and to sample a uniform variate $u^{(r)} \leftarrow$Uniform$(0,1)$ at each sample r. Then we accept if $u^{(r)} < \frac{p(x^{(r)})}{Mq(x^{(r)})}$. For the rectified Gaussian, this just means we use a regular Gaussian to propose, rejecting all samples that are less than zero, and accepting with probability one otherwise.

For $\mathbf{h}$, in order to use rejection sampling, it must be sampled pixelwise, which means the

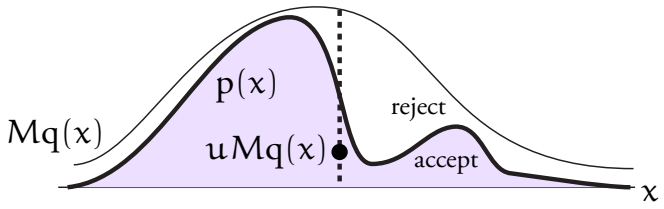


Figure 10.2: *Principle of rejection sampling*

more efficient CGLS solver is not possible, however this is less of a problem due to its smaller dimension than the image (we could also consider a more complex block updating proposed in [214], but we just use the pixelwise scheme here). This requires the PDF of one pixel in the PSF, $\mathbf{h}_{[i]}$, conditional on the rest, $\mathbf{h}_{\setminus i}$:

$$p\left(\mathbf{h}_{[i]} \,|\, \mathbf{h}_{\setminus i}\right) \propto \mathcal{N}^{+}\left(\mathbf{h}_{[i]} \,\middle|\, \mathbb{E}^{k}\left[\mathbf{h}\right]_{[i]} - \mathbf{N}_{[i,i]}^{-1}\mathbf{N}_{[i,\setminus i]}(\mathbf{h}_{\setminus i} - \mathbb{E}^{k}\left[\mathbf{h}\right]_{\setminus i}), \mathbf{N}_{[i,i]}^{-1}\right) \tag{10.2}$$

where $\mathbf{N} = \left(\mathrm{cov}^{k}\left[\mathbf{h}\right]\right)^{-1}$, and the square brackets extract a sub-vector or entry of the matrix. A random update order is used to update the pixels, for $i = 1 \cdots L_{h}$. In practice, we first perform a step of CGLS without $\boldsymbol{\nu}$ (as in the VB solution) to find $\mathbb{E}^{k}\left[\mathbf{h}\right]$, and then sample each pixel once using (10.2). We then resample pixels which became negative, conditional on the new values of the others, until we get a positive sample. We also have to approximate the terms $\mathbf{N}_{[i,i]}^{-1}$ and $\mathbf{N}_{[i,\setminus i]}$. Note that we can calculate $\mathbf{N}$ without inversion, however even finding $\mathbf{F}^{T}\mathbf{F}$ which it depends on is expensive, so we use a BTTB approximation of this matrix. We then work out the rows of $\mathbf{N}$ as required. $\mathbf{N}_{[i,i]}$ is just the i^{th} entry into each row.

10.2.3 Simulated Annealing

The pixel-wise sampling of $\mathbf{h}$ results in slower convergence, or mixing of the Markov chain. simulated annealing (SA) [76, 160] is a popular method to accelerate the search for high probability regions of the parameter space. The posterior distribution is raised to a power $1/\tau$, where $\tau(k)$ is a temperature parameter, such that $p^{(k)}\left(\boldsymbol{\Theta} \,|\, \mathbf{g}\right) = \left(p\left(\boldsymbol{\Theta} \,|\, \mathbf{g}\right)\right)^{1/\tau(k)}$. High temperatures result in the PDFs being less "peaky" and more uniform, avoiding the search getting trapped in local optima. Samples are gathered after burn-in once $\tau(k) \approx 1$. The distribution should be "cooled" sufficiently slowly, in analogy to annealing in metallurgy. We use a quasi-geometric annealing schedule: $\tau(k) = 1 + (\tau(0) - 1) \cdot \kappa^{k}$. The modification to the conditionals using the temperature parameter involves scaling the variance of the Gaussian distributions at iteration k by $\tau(k)$, whereas for the IG PDFs, we make the substitution $\alpha_{\theta} \rightarrow$

$\frac{\alpha_\theta+1}{\tau(k)} - 1$ and $\beta_\theta \rightarrow \beta_\theta/\tau(k)$. We constrain the maximum temperature for these distributions such that $\alpha_\theta > 1$, otherwise the samples will become unbounded.

Once the samples have been obtained, expectations may then be found to provide the desired parameter estimates, for example the MMSE estimate, $\mathbb{E}[\mathbf{f}|\mathbf{g}] \approx \frac{1}{K-k_0}\sum_{k=k_0}^{K}\mathbf{f}^{(k)}$, where k_0 is chosen to allow for convergence during the burn-in period. Other possibilities for inference with the samples include computing a MAP estimate, by cooling the temperature to zero, or to learn more about the posterior by looking at other statistics of the samples.

10.3 Results

We now present several results using the proposed method. In the first experiment (Exp. 1), the Lena image is synthetically blurred with a simulated out-of-focus blur of diameter 12 pixels, and WGN was added to give a BSNR of 40dB. The blurred image in and restored images are shown in fig. 10.3. It is also worth referring to the movie clips of the samples for each experiment (see Appendix H). Note that the samples appear much noisier than the results, because when we calculate the MMSE estimate the samples are averaged together and this removes the noise. In Exp. 2 and Exp. 3 we use the same blur but 30dB and 20dB BSNR respectively. We also show the non-blind restorations using the same algorithm in this case — the results are very similar. Note that the 20dB result in particular ends up shifted by a couple of pixels in the blind restoration; this is very evident when viewing the video (see Appendix H).

In all Exp. 1–3, the observed image support $\mathcal{S}_g$ was 183×183; the PSF support $\mathcal{S}_h$ was 37×37, giving $\mathcal{S}_f = 219 \times 219$; the image and blur block sizes were $m_r = 8$ and $m_{r_h} = 3$ pixels, and the AR orders $P = 5$ and $P_h = 3$. For Exp. 1–3, the annealing used $\tau(0) = 2$ and $\kappa = 0.998$, over $K = 4000$ iterations, with $k_0 = 3000$. Due to the hierarchical model, the algorithm is fairly robust to different choices of the hyperparameters, although reasonable values will aid more rapid convergence. We used the same values as before, but with $\delta_{\alpha_f} = 1 \times 10^{-3}$ to avoid instability of the AR filters. Due to the new segmentations at each iteration, lines 6 and 12 of Algorithm 2 were also repeated twice per main iteration to update the parameters more effectively.

In Exp. 1 and Exp. 2, $\mathbf{h}(0)$ was chosen as a circular PSF of radius 10. This is quite close to the starting condition, but far enough away that it causes problems if we use the VB method. If we

(a) $\mathbf{g}$, simulated (b) $\hat{\mathbf{f}}$, ISNR=6.31dB (c) $\mathbf{f}$ (region $\mathcal{S}_g$ shown)

(d) $\mathbf{g}$, simulated (e) $\hat{\mathbf{f}}$ (f) $\hat{\mathbf{f}}$, non-blind

(g) $\mathbf{g}$, simulated (h) $\hat{\mathbf{f}}$ (i) $\hat{\mathbf{f}}$, non-blind

(j) $\mathbf{g}$, photographed (k) $\hat{\mathbf{f}}$, ISNR=6.77dB (l) $\mathbf{f}$ (region $\mathcal{S}_g$ shown)

Figure 10.3: *MCMC results: (a) – (c) Exp. 1; (d) – (f) Exp. 2; (g) – (i) Exp. 3; (j) – (l) Exp. 4;*

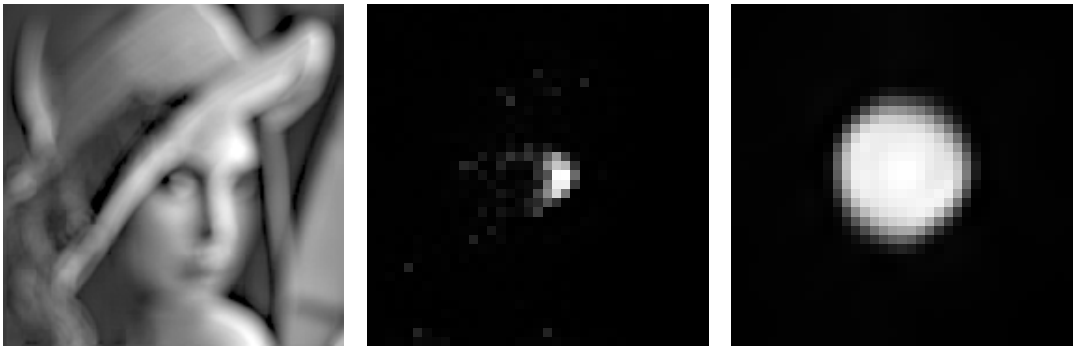
were to use a multiscale approach it is also likely we would have an initial estimate at least this close. In the lower BSNR of Exp. 3 we found convergence was generally faster, and we started with a Gaussian PSF of standard deviation 2.5. In other experiments at higher BSNR we use a more distant starting condition, but with slower annealing to ensure proper exploration of the space.

10.3.1 Real Blurred Image

In Exp. 4, the image was degraded by a real camera. The Lena image was displayed on a computer screen and photographed with the camera lens out-of-focus to produce a similar blurred image to that in Exp. 1 (in fact it is slightly more blurred). The green channel of the image only was used; this was bilinearly interpolated and downsampled, in order to avoid demosaicing. It is also important to work in a linear colour space, and convert to a gamma 2.2 space for final display. Despite the care taken to ensure a linear image, there is undoubtedly some residual non-linearity in the process, as can be seen from the results in fig. 10.3. The same parameters were used as in Exp. 1, although the annealing was modified to $\tau(0) = 3, \kappa = 0.999, K = 10000, k_0 = 9000$ to ensure better convergence. It is also possible that the PSF is slightly spatially varying in this case, as more artefacts are visible away from the centre.

In order to calculate an ISNR value, another image shot under the same conditions but with the lens in focus was used as a reference (fig. 10.3(l)). This was aligned with the restored image and the images histogram matched, which avoids the shift and scale ambiguities and enables calculation of the metric.

In fig. 10.4 the result obtained with the VB algorithm using the same AR prior as described in Chapter 9 is shown for comparison. 4000 VB iterations were used, beyond which no significant change was seen. The same initial conditons are used as in the MCMC case - a 10 pixel diameter circular blur, which is close to but slightly smaller than the true blur. It is clear that although some slight improvement has been produced, the PSF tends towards a delta function and correct convergence is not obtained in this case, unlike with fig. 10.3(k).



(a) $\hat{\mathbf{f}}$, VB method (b) $\hat{\mathbf{h}}$, VB method (c) $\hat{\mathbf{h}}$, MCMC method

Figure 10.4: *(a) Unsuccessful blind restoration of fig. 10.3(j) using VB method and AR prior from Chapter 9; (b) Corresponding estimated PSF; (c) Estimated PSF corresponding to fig. 10.3(k).*

10.3.2 Different initial conditions and PSFs

In Exp. 5 (not shown in the figure, but see Appendix H), we use the same PSF as Exp. 1, but we initialise with a variance 3 pixels Gaussian, which is a lot further from the true PSF. In order to converge to the correct region of the posterior, we find we need a slower annealing using $\kappa = 0.9993$ with a higher starting temperature of $\tau(0) = 4$. We can see by referring to the video clips in Appendix H that in general, at higher temperatures the image resembles more noise, and as it is cooled, the edges begin to appear, followed by lower frequency content. It seems critical that the cooling is sufficiently slow that the edges are more or less correct before it is cooled to a lower temperature that starts to resemble the proper image.

The remaining experiments are shown in fig. 10.6. In Exp. 6, we use a synthetic 1D motion blur, of 11 pixels, at 40dB BSNR, with the same initialisation and settings as Exp. 5. In Exp 7. we show the Lena image, blurred with a 7×7 square PSF at 45dB BSNR (this is the same experimental conditions which are used in [217]). In Exp 8. we use the same settings but with the Cameraman image. Note that all of these cases, it is obvious from considering the restored PSF that there is a sub-pixel shift in the restoration.

Finally, in Exp. 9, we use a "hand drawn blur" (at 40dB BSNR) which is supposed to represent an arbitrary camera motion, and represents a challenge for any BID algorithm (see fig. 10.5). Nevertheless, we were able to get a good result by making one small change to the algorithm. We noted that using a lower value of $\delta_{\alpha_f} = 1 \times 10^{-5}$ helped to give a sparser image, where the edges converged much more rapidly, although the low frequencies never returned. We

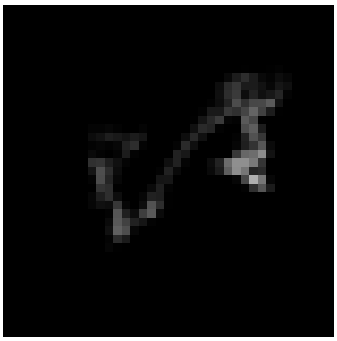


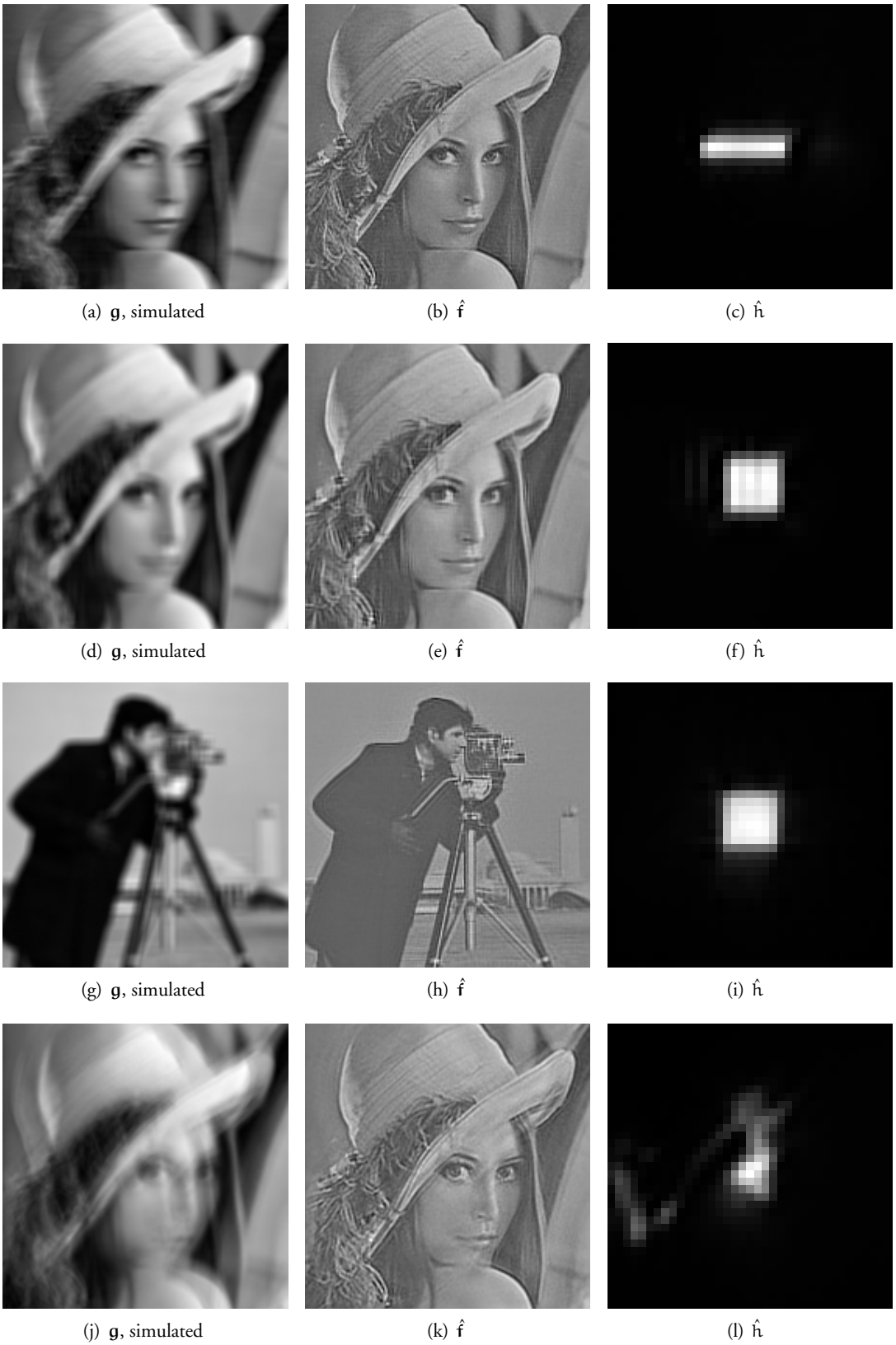
Figure 10.5: *Hand drawn PSF, to represent motion blur*

then tried slowly raising the value of δ_{a_f} back to 1×10^{-3} at the end of the annealing period, and we obtained a reasonably good result. Note that the PSF in this case contains lots of small detail, and we have recovered most of it, albeit with a substantial shift. Because of the Gaussian initialisation, it seems that we tend to converge to a solution where the maximum intensity of the PSF is located at the origin. It is interesting to note that although we have the shift ambiguity, it does not seem to adversely affect the restorations to a significant degree. It does mean that they can drift very slowly over a long period of samples, and this will lead to a MMSE estimate that is not as sharp as it could be. However, as long as we keep our averaging to a couple of thousand samples, it doesn't seem to be a problem.

10.4 Conclusions

In this chapter, we have shown that MCMC methods can be used in BID to tackle the very challenging optimisation problems set by these models, as well as to gain insight into the quality of the obtained solutions. Very flexible and locally adaptive image and blur models have been proposed, that are capable of modelling a variety of types of blur, and adapting to local features and textures in the image. We have shown that the method is capable of visually very impressive results that we hope make a useful step forward in quality in the field of BID.

One common criticism of MCMC methods is with speed. It is true that the simulations presented here are not currently the fastest in the world, taking around 20 to 30 seconds per main iteration (i.e. typically we need a couple of days to run the complete experiment). However for the quality of the results, sometimes things are worth waiting for. If the photo is from a one-off event then the waiting time may not be important. Improvements in central processing unit (CPU) performance will bring down run times, although it is worth considering new



(a) $\mathbf{g}$, simulated (b) $\hat{\mathbf{f}}$ (c) $\hat{\mathbf{h}}$

(d) $\mathbf{g}$, simulated (e) $\hat{\mathbf{f}}$ (f) $\hat{\mathbf{h}}$

(g) $\mathbf{g}$, simulated (h) $\hat{\mathbf{f}}$ (i) $\hat{\mathbf{h}}$

(j) $\mathbf{g}$, simulated (k) $\hat{\mathbf{f}}$ (l) $\hat{\mathbf{h}}$

Figure 10.6: *MCMC results: (a) – (c) Exp. 6; (d) – (f) Exp. 7; (g) – (i) Exp. 8; (j) – (l) Exp. 9;*

architectures such as parallelisation and general purpose graphics processing unit (GPGPU) acceleration, in addition to algorithmic improvements.

Possibilities for future work will include better modelling of the AR parameter distributions, imposing positivity on the image, and achieving more rapid convergence of the Markov chain. One obvious possibility for speed-up is to use a multi-scale approach, beginning with a coarse grid and working to the final resolution. Methods such as preconditioning can help accelerate the CG algorithm, or alternative numerical solvers can be investigated. Finally, other methods from the rich MCMC literature can be applied to speed up searching of the parameter space.

Chapter 11
Conclusions and Future Work

11.1 Conclusions

In the first part of this thesis, the BID problem has been formulated, and the problems that are encountered and the models and methodologies that can be used to solve the problem were considered. In Chapter 2, the 2D convolution model and PSF formulation of the observation model was thoroughly examined. In Chapter 3, the classical regularisation approach to image deconvolution, and some restoration techniques were described. In Chapter 4, the Bayesian paradigm was introduced, as a robust and principled way for modelling and solving inverse problems such as BID; some existing blur and image models that are used as priors were then reviewed. A review of Bayesian solutions to BID with a survey of existing methods in the literature was given in Chapter 5.

In the second part of the thesis, the novel image and blur prior models were developed in Chapter 6, before considering the 3 proposed Bayesian methods for estimation of the unknown image and blur in the following chapters. The marginalisation of the image and other model parameters was considered in Chapter 7, in order to derive a closed form expression that could be used in an optimisation framework to estimate the unknown PSF, in the MMAP method. While this method provided useful insights into the problem, unfortunately it is not generalisable to a wide class of standard PSFs because of problems with causality. Therefore it was proposed to study the full model (including simultaneous image estimation) which does not suffer from these problems. However as this full model does not have an analytic solution for maximisation or finding expected values, we had to propose either to approximate the distribution in Chapter 9, or simulate it in Chapter 10. On the way we had to consider how to deal with boundary effects in order to deal with the required image restoration steps correctly in Chapter 8. In Chapter 9, we used the VB approach to approximate the posterior and estimate both the image and blur along with the unknown model parameters. This showed some encouraging results in the case of Gaussian blur. However for other types of blur, we had to resort to using full simulation of the posterior distribution using MCMC methods

in Chapter 10. Although relatively computationally intensive, these methods allow for the highest quality restorations, and we have shown blind restorations that are comparable to non-blind results, with both synthetic and real data, and with a variety of different types of PSFs.

11.2 Suggestions for Future Research

11.2.1 New image priors

Although new priors for image (and blur) modelling have been proposed in this work, there are many competing models which have been recently published, particularly in areas such as image denoising, inpainting, and demosaicing, as well as SR and non-blind deconvolution. It would obviously be worth investigating the differences between some of these and the proposed models, as well as whether any of the features of these models can be adapted for use in the BID problem.

For instance, multiresolution models have also been proposed to represent the interactions between pixels at different scales. Wavelet models attempt to represent the image in terms of a filterbank that transforms the image into a relatively sparse representation at different scales; methods such as Gaussian scale mixtures (GSMs) [177] attempt to model the statistical relations between and across scales, matching empirical distributions. Sparse representations using trained dictionaries [64], or example-based database methods [65] are also used, whereby a collection of typical image patches is built up from either a training image, or from the image to be restored itself. Overcomplete dictionaries are typically used in image restoration; these methods may be seen as a generalisation of standard orthogonal wavelet transform methods, where rather than assuming a known filterbank, typically a solution is sought in which the image is represented as the best sparse approximation of patches from the dictionary. Other training based methods have learned MRF model parameters, for instance with the fields of experts (FoE) models [190].

Both multiscale and dictionary learning based methods could be applied as an extension of the nonstationary AR models, to model texture at multiple scales, and to better represent the likely textures that are encountered in an image. Another possible model that could avoid some of the problems with feedback and "hot pixels" encountered in the VB solutions would be to model the spatial interactions between the image variances, also with a type of MRF model where the variables are constrained to be positive.

11.2.2 New blur priors & Spatially Varying blur

Separate modelling of the PSF is another issue that has received very little attention. By building better parametric or other statistical models which more accurately represent the different types of degradation that are encountered, we could hope to solve the problem a lot more robustly. The example based approach may be one way to do this without using explicit parametric models, but by training a dictionary on actual PSFs from real blurred images.

In addition, the SV blur problem must really begin to be addressed in earnest in order to make these BID methods more suitable for application to real blurred photos. Real blurred images, whether due to optical aberrations, defocus, or motion of the camera typically suffer from SV blurs, different in each region of the image. The extension to their estimation would require either the extra information from a multiple observations, or an extended parametric hierarchical model for the blur to add some spatial consistency or reduce the parameter space. In the multichannel case, either mild continuity constraints such as a MRF depth-map, or a simple sectioning based approach should help. The challenging single channel SV case has fewer constraints to help infer the blur, and so strong blur priors are required in order to facilitate their estimation.

The case of space-varying motion blur in particular is a particularly challenging but potentially rewarding problem to consider solving, as this is one of the most common blurs in practice, and perhaps where the most work is still to be done in modelling. In practice we often have an arbitrary motion of the scene including time-varying rotation and translation during the exposure time. This PSF will be spatially varying, yet if we consider rigid body motions, then there will be a complicated functional relationship between the PSFs in each part of the scene which could be used as part of the image recovery procedure. It is interesting to note, that despite the complicated degradation, these types of blurs are one of the most well-posed due to the high frequencies typically present in the random-walk type PSFs, so in theory we could recover much information from these images.

11.2.3 Acceleration of algorithms and deployment on new hardware

Many of the methods used rely on solving optimisation problems, for instance using CG algorithms, or some of the other priors in the literature require nonlinear search procedures over sparse dictionaries. With computationally heavy algorithms of the kind described and scale

necessary for processing the increasingly large images encountered in real scenarios, and the change in the evolution of CPU designs, it is becoming a reality that signal processing methods cannot simply put off implementation details as a secondary step; again thought should be given during the algorithmic design stage. Exploring the parellelisation of the proposed algorithms is one option, for example MCMC techniques such as parallel tempering [62] could be explored. Further hardware acceleration is being proposed through the use of mathematical libraries that allow linear algebra to be performed on readily available graphics processing units (GPUs), giving rise to a trend in general purpose graphics processing unit (GPGPU) computing. These techniques hold great promise to help speed up the algorithms.

11.2.4 Multichannel constraints

Work on the multichannel blind deconvolution (MBD) problem [205, 206] has shown that under certain conditions, the original image and blurs may be reconstructed perfectly in the noiseless case; in the noisy case this formulation provides extra constraints that the unknown blurs must satisfy. The required conditions of *blur-coprimeness*, that is the blurs in each image contains no common factor but a scalar, are generally applicable in practice. The MBD class of problems may be subdivided into single input multiple output (SIMO) and multiple input multiple output (MIMO) problems – the former is essentially connected to the motion-free super-resolution problem while the latter is typically for colour imaging with cross-channel blurs. We could apply these constraints to the proposed models to help to better solve multichannel problems.

11.2.5 Application of proposed models and methods to new areas

Because of the general form of the Bayesian framework we have used, the priors and algorithms that have been developed may also be used with other types of observation models, for example SR, denoising, inpainting, and so on. Methods that combine SR and BD, resulting in the blind superresolution (BSR) problem, have only just begun to receive attention [204], and hence this is a field ripe for investigation. In a similar manner, joint motion estimation and high resolution (HR) image estimation in the SR problem is something that has received only a little attention in the Bayesian framework [176, 212], but application of MCMC could help with this problem.

However it is worth considering extensions beyond traditional image processing areas that just attempt to compensate for defects in existing imaging procedures. In the photographic observation process there are several trade-offs. A standard camera system captures information through a set of lossy processes; a projective transformation, integration and discretisation of the light rays over each position of the sensor, sampling of the electromagnetic spectrum, and integration of camera/scene motion during the exposure period, as well as corrupting noise processes. We can view the available light information as a multidimensional function of space, rotation, wavelength and time, of which a typical image only captures a small subset. This function has been termed the *Plenoptic function* [2]. Arguably, by capturing a different subset we may result in a different set of data that may be more appropriate to our end cause, though we may require intermediate processing to realise this result.

By considering the possibilities of novel imaging modalities, further gains may be made. That is to say, rather than constraining attention to restoration of existing images, where signal processing can only be applied as an afterthought, one may ask how the observation system could be designed in conjunction with the restoration stage to give an overall design optimal for purpose.

11.2.6 New imaging methods

Recently, a number of new methods for capturing natural images have been proposed, and are rapidly appearing in commercial hardware. Methods such as wavefront coding [44] have suggested modifying existing camera lenses in a manner that provides a special pre-capture distortion to the image. This distortion is then removed in software, resulting in an image that is in focus at all different depths, but whilst maintaining the high SNR available from large-aperture lenses. Another technique uses a hybrid stills/video camera to get optimal spatiotemporal resolution by recording the motion path and using this to deblur the still image [15].

Another novel design is that of the "Plenoptic Camera" [2, 165]. Rather than using just one lens to focus the image onto the image plane, this uses an additional array of microlenses, somewhat like a fly's eye, to give a set of images from slightly different perspectives, which has enabled synthetic refocussing, and could allow multichannel deconvolution or SR to be performed as postprocessing steps — work along these lines in superresolving a simulated compound eye system has been proposed recently [53], and for application with respect to iris

recognition in [13].

Other ideas using coded or programmable apertures to eliminate motion blur [4, 183] or enable depth estimation for guided deconvolution [130, 131, 219] have also been produced, and could benefit from further development of reconstruction algorithms.

Collectively, these developments have been termed *Computational Photography*, or *Computational Imaging*. They represent a general trend in imaging systems, describing the convergence of techniques in computer vision, computer graphics, opto-electronics and signal processing [135]. All these examples show that information may be captured in novel ways, and that there is a need to develop and improve existing signal processing algorithms to benefit most from the available data. It is hoped with alternative imaging strategies, combined with increasingly cheap and abundant computational power for signal processing, future cameras and imaging devices will be able to be produced significantly more cheaply than is currently possible.

A related research direction is *compressed sensing*. A team at Rice University have recently made a prominent announcement of developing a "Single Pixel Camera" [221], which utilises this method to actually capture a compressed image by using projections onto random grid patterns; reconstruction algorithms may be used to recover the original image. Among other applications, it has been suggested that this device could be used to make infra-red cameras two orders of magnitude cheaper than traditionally. The new theory underlying these methods may usher in a new era in signal processing whereby the standard tenets of the Nyquist limit for sampling bandlimited signals may be avoided. This theory suggests that if the signal is compressible using a sparse representation, then there is no point trying to capture the samples in a regular manner; we are wasting information where there is little present, and discontinuities represent an impassible barrier, unless we capture in a manner that is not so wasteful of information.

Clearly there are close links to image restoration in all of this work, and new restoration methods should be designed with these new types of hardware in mind.

Part III

Appendices

Appendix A
Camera Models

In this appendix, we begin by looking at how an ideal camera would capture the image, in order to give an idea of what we are trying to restore. Then we consider various effects of real cameras that mean the captured image deviates from the ideal image. Some of these models represent approximations of the true degradations, but are necessary to begin to analyse the problem. We will look which of these effects can influence the commonly used *point-spread function (PSF)* model (the 2D impulse response) for the blur. In the approaches presented in this thesis to solving the BID problem, we will use the general spatially invariant PSF model and not explicitly use the camera models; however it is important to be aware what affects this, and why many overly simple parametric models may ignore important real aspects of the degradation process.

A.1 Pinhole Camera and the "true" image

In this section, we consider a physical camera model representing the observation process, including various non-ideal effects that may need to be compensated for. We begin by giving a definition of the true image that will be used for our purposes as a goal for restoration.

Consider a three-dimensional scene we wish to image. Light is emitted, transmitted, or reflected, from a continuum of points in the scene. The most exact approach to describe how the light propagates through space and interacts with objects would be to consider it as an electromagnetic field and solve Maxwell's equations for the region of interest [29]. This is often either unfeasible or overly complicated, and instead a geometrical model is more often used as a starting point with further refinement as necessary. The geometrical model is particularly appropriate for a simple analysis of optical systems; commonly the wavelength of light is much smaller than objects it passes, and diffraction effects may be ignored.

A related representation of the light propagation that is useful for developing imaging models is the *light field* or *plenoptic function* [1], a vector field in free space. At each point in space this

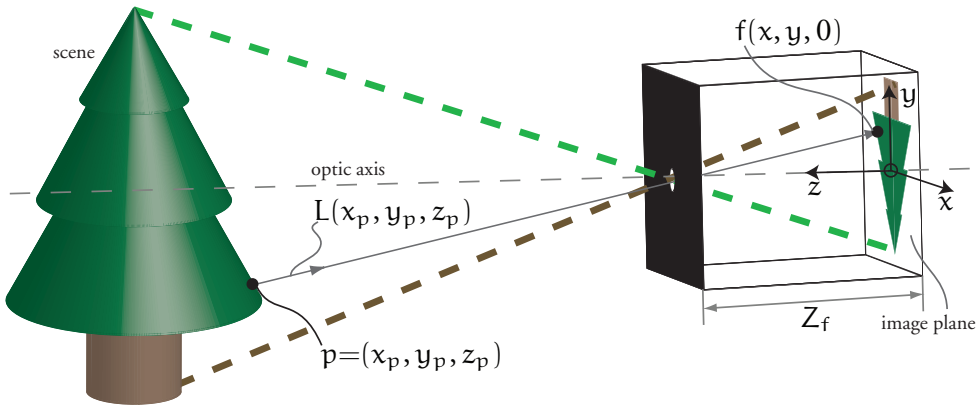


Figure A.1: *Pinhole camera model*

vector will represent the spectral distribution of light according to direction and time. In this general form the plenoptic function is a 7-dimensional quantity: $L(x, y, z, \theta, \phi, \lambda, t)$, where $\{x, y, z\}$ denotes position, λ wavelength, $\{\theta, \phi\}$ angle and t time. Simplified versions of this notation consider only a subset of the variables, for instance only monochromatic light, and a single time instant. If the scene is assumed to be *Lambertian*, with light reflected uniformly in all directions from each surface, then we can ignore the the angular coordinates. Even though we will only considering restore monochromatic images (not videos) in our final representation, it is important to realise these extra dimensions can play a role in the degradation process, through for instance motion blur and chromatic aberration.

To form a "perfect" or "true" image of the scene, assume for now that the light travels out as rays through the homogeneous medium, in straight lines from their source (which could be a reflection from an object) until they intersect with an imaging plane. The *pinhole camera* model (fig. A.1) presents an infinitesimal aperture between the scene and the imaging plane, such that each point in the scene maps to a single point on the imaging plane. The continuous function $f(x, y, 0)$ represents a perspective projection onto the 2D imaging plane at $z = 0$ by discarding the depth information in the following manner:

$$L(x_p, y_p, z_p) \rightarrow f(x, y, 0) = f\left(\frac{Z_f x_p}{z_p - Z_f}, \frac{Z_f y_p}{z_p - Z_f}, 0\right) \tag{A.1}$$

where $L(x_p, y_p, z_p)$ represents the light field intensity or radiance emitted from a point p in the scene, in the direction passing through the pinhole, which is located at $\{0, 0, Z_f\}$ on the optical axis; Z_f is the focal length. Each point in the scene that is visible to the pinhole has a

unique image point on the image plane. The resulting light field, from all such points, that passes through the pinhole and impinges on the imaging plane is the continuous version of the *true* image, denoted $f(x, y)$. In practice we will try to restore an optimally anti-aliased and sampled version, $f(\mathbf{s})$, of this image; equivalently it may be assumed that the original scene is band-limited.

In order to connect the ideal camera model with the actual linear observation model and the vector-matrix formulation that will be employed, we can imagine that the true image $f(x, y)$ actually appears as the input signal to an imperfect camera system.[1] In reality this idealised representation of the true image is never recoverable exactly, however it provides a useful target to aim for in our restorations.

A.2 Terminology: Enhancement or Restoration?

It is worth reflecting again on the notion of the "perfect" or "true" image and the terminology used. From the Bayesian point of view, the objective of *restoration* is to try and infer an improved representation of the underlying scene, in a probabilistic sense, based on the available information, rather than obtaining a particular true restored image. The term *image enhancement* is often used in image processing to refer to procedures that attempt to improve the perceived quality of an image without necessarily modelling the degradation process. Typically the term deconvolution can only be used when solving a problem where the linear observation model strictly applies, which is usually an approximation. However, as the goals are similar — to recover a better quality image — the terms image restoration, enhancement, and deconvolution will be used here more or less interchangeably.

A similar argument may be applied to the PSF: although the idea that the blur constitutes the exact image of a point source that has passed through the system under consideration, in practice the PSF, $\mathbf{h}$, and the convolution that will be implied in (2.6) is merely a convenient mathematical tool for representing the problem. In reality, the lens in the camera model provides a highly complicated non-linear transformation of the lightfield to give the observation, of which the geometric and Gaussian optics models give a first order approximation, ignoring the wavelength of light and other effects [29]. However, as it is usually just the restored image that we are interested in recovering, these concerns need not raise too much concern

[1] Assuming the scene is planar, or the PSF is independent of depth, otherwise we should also estimate a depth map.

in practice, and the PSF model is a convenient tool for mathematical analysis. Blind restoration enables a variety of physical effects to be absorbed into the PSF term and thus provides resilience against poor modelling.

Some of the effects that are present in real world camera systems will now be analysed and described; an overview of modelling some of these effects was given in [129]. Following this, the most important effects that will be studied for the observation models used in this thesis will be detailed.

A.3 Finite exposure time & Motion of the Camera or scene

A real camera has a shutter which opens to let light in for a certain duration; with the ideal camera model it is assumed this exposure time is instantaneous. This would require the intensity of the light field to be infinite, or the sensor to be infinitely sensitive, in order to collect any light, so in practice the shutter is opened for a finite time until sufficient light is gathered to properly expose the image. Observe that any relative motion between the camera or scene during this time will result in an integration of the time-varying light field, appearing as motion blur in the image. In practice unless the motion is constrained to be planar to the image plane at a constant depth, we would need to use a SV motion model. This makes this type of blur very difficult to restore in practice, and typically it is assumed the blur can be approximated as locally constant in order to do any deconvolution. Proper modelling of camera motion and its effects on motion blur should be the topic of further research.

A.4 Camera with lens: Gaussian Optics model

Rather than the pinhole model where the aperture is vanishingly small, most practical cameras employ a larger aperture, or diaphragm, with a series of lenses to collect a larger amount of light, whilst still forming a focused image at the image plane. This enables shorter shutter speeds to be used to help freeze motion. Following the geometrical model, in order to analyse the resulting light field that now appears on the image plane, the best analysis is obtained via ray tracing. That is, each ray $\mathrm{L}(\mathrm{x}_\mathrm{p}, \mathrm{y}_\mathrm{p}, \mathrm{z}_\mathrm{p}, \theta_\mathrm{p}, \phi_\mathrm{p})$ that is emitted from the set of points visible to the camera aperture (or entrance pupil) is traced through the optical system, refracting at

the air-glass boundaries according to Snell's law [29] (see fig. A.2)

$$n_{air} \sin \theta_i = n_{glass} \sin \theta_r. \tag{A.2}$$

The ray tracing principle may be used to analyse the focussing effect of a lens. Although it is not the ideal shape, lenses commonly have spherically shaped surfaces due to ease of manufacture. Applying (A.2) to a single spherical refracting surface, along with simple geometry considering the two triangles shown in fig. A.2 gives the relation (see [156]):

$$\frac{U + R}{W - R} = \frac{n_{glass} \sin \theta_w}{n_{air} \sin \theta_u} \tag{A.3}$$

Here R is the radius of curvature, and U and W the distance from the object point to the image point on the optical axis. The distance to the image point clearly depends on the angle of the original ray.

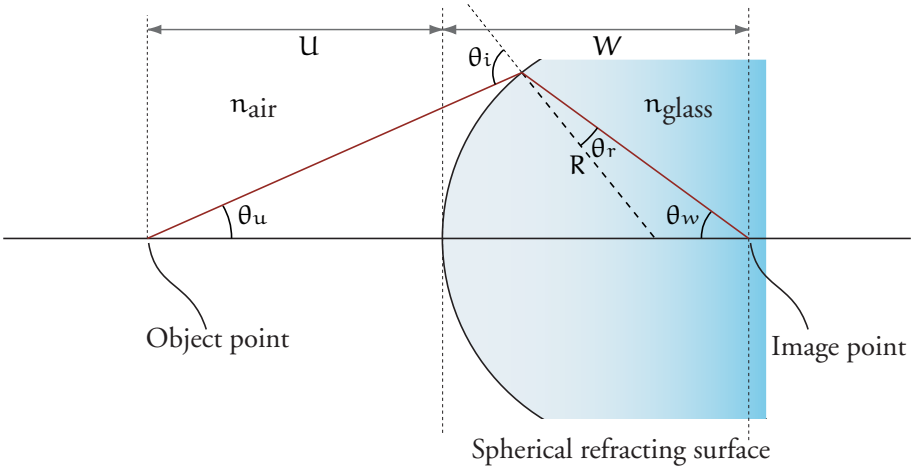


Figure A.2: *Effect of a spherically shaped refracting surface on an incident ray*

For a simplified analytical analysis, the Gaussian optics model is often used. In this model, it is assumed that objects in the scene are sufficiently far from the lens that the rays make very shallow angles with the optical axis. Then the *paraxial* approximation is employed such that:

$$\sin \theta \approx \theta \qquad \cos \theta \approx 1 \qquad \tan \theta \approx \theta \tag{A.4}$$

With the paraxial approximation it becomes simple to analyse refraction effects caused by a simple lens. Equation (A.3) then simplifies, using $\frac{\sin\theta_u}{\sin\theta_w} = \frac{W}{U}$, after some rearrangement, to

$$\frac{n_{glass}}{W} + \frac{n_{air}}{U} = \frac{n_{glass} - n_{air}}{R} \tag{A.5}$$

If the radius of curvature is small compared to the thickness of the lens, the *thin lens* approximation may be used, which simplifies analysis further. With these assumptions, we can find the focussing effect of a thin lens on an incident ray, by applying (A.5) twice, once for each surface of the lens, and taking n_{air} as 1. We then obtain the thin lens equation:

$$\frac{1}{U} + \frac{1}{V} = \left(n_{glass} - 1\right)\left(\frac{1}{R_1} - \frac{1}{R_2}\right), \tag{A.6}$$

where R_1 and R_2 are the radii of curvature of the two surfaces, and V is the distance from the lens to the image point from the second surface. Defining the *focal length* of the lens, Z_f as the reciprocal of the RHS, results in

$$\frac{1}{U} + \frac{1}{V} = \frac{1}{Z_f}. \tag{A.7}$$

Thick lenses, or more complicated systems of lenses may also be handled by finding a thin-lens equivalent system. With this representation, two *principal planes* are identified such that the thin-lens model holds for rays entering and leaving the system; a black-box approach is used so that details of what happens between the principal planes are not required. Analysis of a system of lenses is frequently done using a matrix to represent the linear transformation applied to the position and angle of rays in the incident and refracted lightfields, assuming the paraxial approximation.

A.5 Depth of Field under Gaussian Optics model

Given the assumptions described in the previous section, points in the object at the plane of focus are sharply mapped to points on the image plane. We would like to find out how the points at other depths are affected by the optical system. The thin lens model with a finite aperture is shown in A.3. Taking the centre of the lens as the origin, the cone, or "pencil" of rays emanating from the point in focus at $-U_0$ on the axis is refracted by the lens and focused at the point V_0. These points are related by (A.7), as $\frac{1}{U_0} + \frac{1}{V_0} = \frac{1}{Z_f}$. Similarly, the pencil of

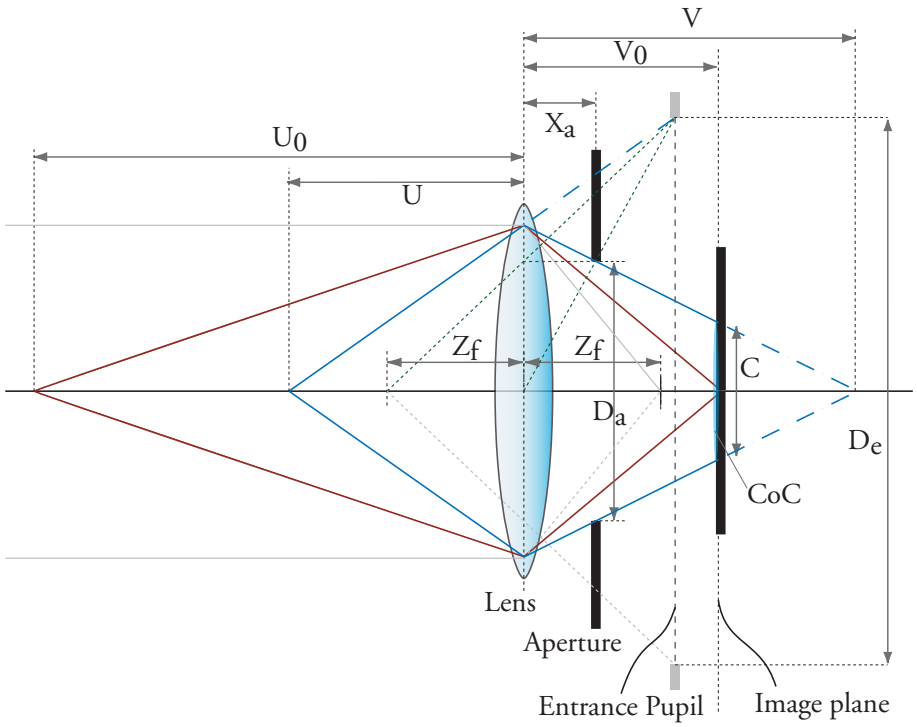


Figure A.3: *Model of Camera with Defocussing*

rays at $-U$ converges after passing through the optical system, forming an image point at V. However the image plane is located at V_0; the intersection of the pencil of rays and the plane is circular in shape. This is termed the CoC, and describes the PSF under this model for a point that is out of focus.

The diameter of the CoC, C can be found as a function of the depth of the out-of-focus point. It also depends on the depth of the in focus point, the focal length of the lens, and the size or position of the *entrance pupil*. The entrance pupil is the (possibly virtual) image of the *aperture stop* seen from the object side. The aperture stop is the diaphragm or lens element that most limits the angle of the rays entering from the point of interest.[2] The diameter of the aperture itself, D_a, is proportional in diameter to the entrance pupil, D_e, according to:

$$D_a = \frac{Z_f - X_a}{Z_f} D_e. \tag{A.8}$$

This is found by similar triangles and application of the lens law. It is common to state the

[2]Note that it is possible for different points of interest to have a different aperture stop, for instance the diaphragm at X_a is the aperture stop for the rays coming from $-U$, whilst the lens itself is the aperture stop for the rays emanating from $-V$.

brightness of the lens in terms of the relative aperture, or f-stop, F, as the ratio of the focal length to the entrance pupil diameter, such that $D_e = \frac{Z_f}{F}$.

Now by again considering similar triangles in fig. A.3, it is seen that

$$\frac{D_a}{V - X_a} = \frac{C}{|V - V_0|}, \tag{A.9}$$

and by eliminating V and V_0 using (A.7) we obtain:

$$\begin{aligned} C &= \frac{Z_f - X_a}{F} \frac{|V - V_0|}{V - X_a} = \frac{Z_f - X_a}{F} \left| 1 - \frac{V_0 + X_a}{V - X_a} \right| \\ &= \frac{Z_f - X_a}{F} \left| 1 - \frac{Z_f U_0/(U_0 - Z_f) + X_a}{Z_f U/(U - Z_f) - X_a} \right|. \end{aligned} \tag{A.10}$$

This expression may be split into two regions, depending on the relative position of U. Taking the case $X_a = 0$ (aperture at the lens) for simplicity, it reduces to:

$$C = \begin{cases} \frac{Z_f}{F} \left(1 - \frac{U_0(U - Z_f)}{U(U_0 - Z_f)} \right), & Z_f < U < U_0 \\ \frac{Z_f}{F} \left(\frac{U_0(U - Z_f)}{U(U_0 - Z_f)} - 1 \right), & U_0 < U < \infty \end{cases} \tag{A.11}$$

This relation is shown in fig. A.4.Note that the effective resolution limit of the film, or sensor sampling grid, denoted here as δ, effectively means that a finite range of depths, U^- to U^+, appear to be perfectly sharp - this range is known as the depth of field (DoF). Alternatively δ may be based upon the resolving power of the human eye, given the size of the resulting print or display and the viewing distance. The distance in front of U_0 where the defocus is the same as that at infinity is termed the *critictal distance*, K. In practice, lens aberrations and diffraction will mean there is a limit to the achievable resolution; sometimes an equivalent *circle of least confusion* will be given for the smallest radius PSF an optical system can produce.

A.6 Lens aberrations

In order to analyse departures from this model, a power series representation of the trigonometric function is used, i.e.

$$\sin\theta = \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \cdots \quad \cos\theta = 1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \cdots \quad \tan\theta = \theta + \frac{\theta^3}{3} + \frac{2}{15}\theta^5 + \cdots \tag{A.12}$$

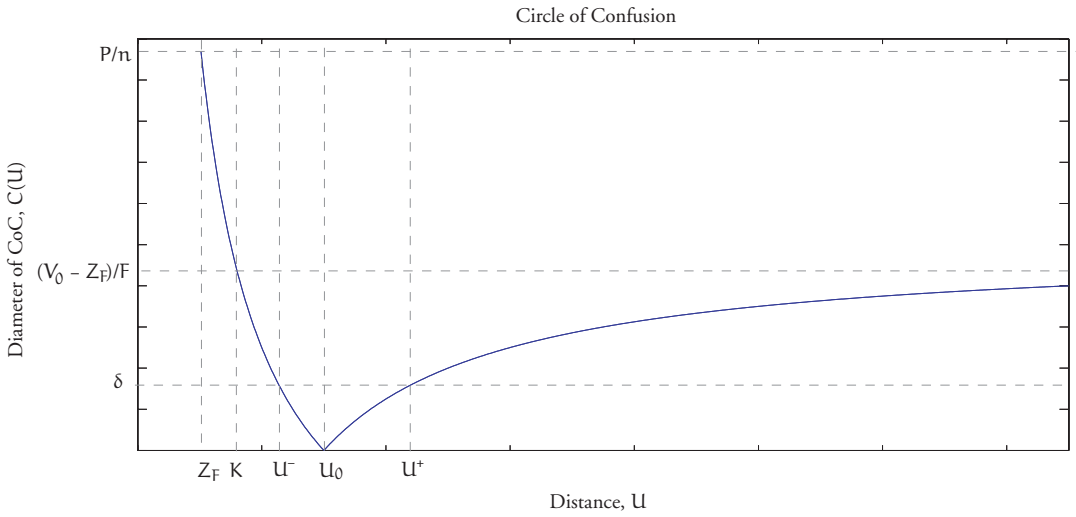


Figure A.4: *Depth of Field and Variation of CoC with scene depth*

Including terms of $O(\theta^3)$, the third order theory, or *Seidel aberrations* become apparent [29, 156]. These aberrations describe the deviation away from the paraxial approximation for the ideal thin spherical lens model. Each is considered assuming the others not to be present. The Seidel aberrations are:

- Spherical aberration;
- Coma;
- Astigmatism and Field Curvature;
- Distortion.

The first order approximation holds for the rays originating close to the optical axis passing through the centre of the lens, but is reasonably accurate up to a deviation of around 5 degrees from the optical axis. For the wider angles of view common in photography, the aberrations therefore play an significant role.

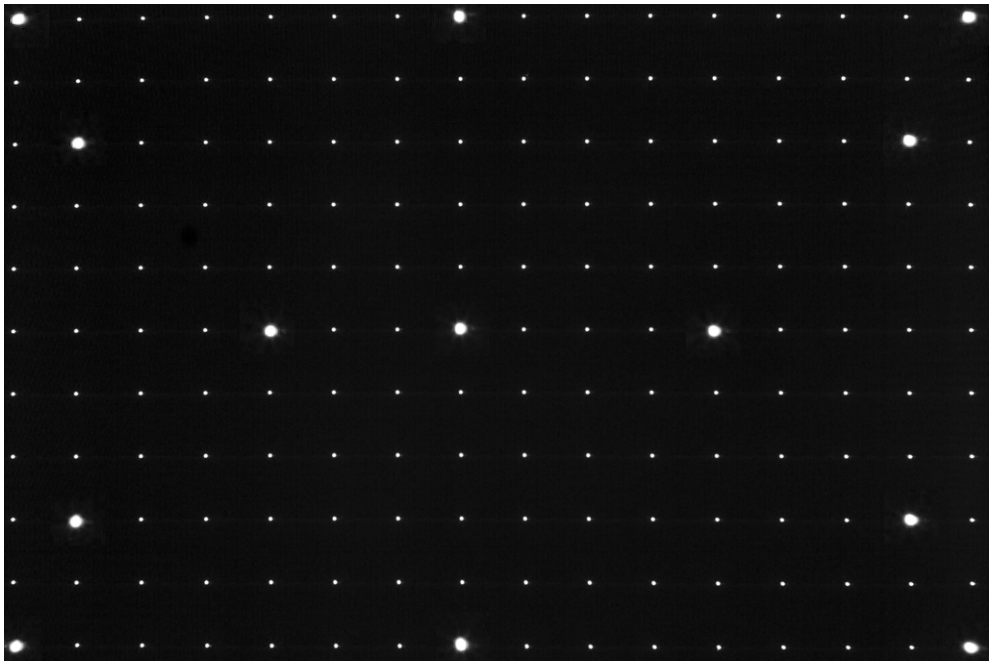
Spherical aberration manifests in all lenses using only spherical elements as an apparent change in focal length for *marginal* rays — those striking the lens at more obtuse angles and away from the optical axis. These rays have a steeper angle of incidence and are refracted more strongly. Thus scene points are not sharply imaged at one plane but at a range of image points in front

of and behind the image plane. The paraxial rays still give a sharp point, but the marginal rays of the same point tend to spread out around this in a halo-like PSF on the image plane.

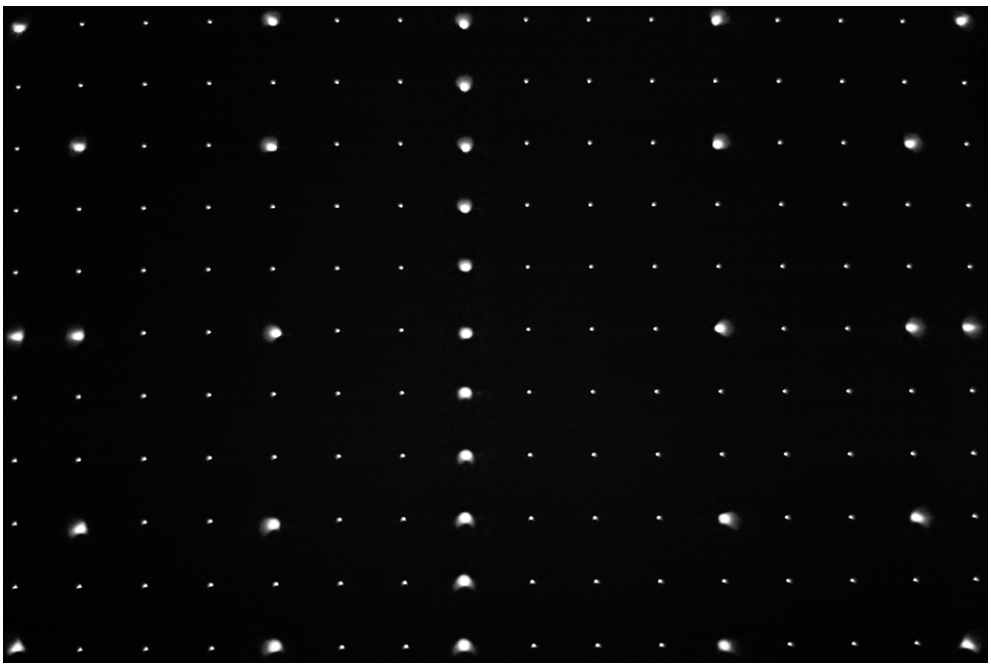
Coma affects imaging of off-axis points in the scene. Whereas the principal or *chief* ray, passing through the centre of the lens (or the aperture stop), hits the image plane at the location expected by the Gaussian model, marginal rays of the same point are focused at different points on the image plane; effectively they have a different magnification. The overall effect, taking account of all rays from this point passing through the lens, is to give a PSF that is a trail of expanding circles rather than a sharp image point. This *comatic flare* is more pronounced for points away from the centre of the image plane [43]. See fig. A.5 for an example.

Astigmatism and Curvature of Field are two related abberations that occur also for off-axis points even for lenses compensated for spherical aberration and coma. The marginal rays coming from an off-axis point will be refracted by different amounts and hence focus at different depths in front of or behind the image point of the chief ray. The rays in the *tangential* plane, defined by the object point and the optical axis, will focus at a different depth from those in the *saggital* plane, which is perpendicular to this. The result is that elliptical PSFs will be present, termed astigmatism. Their shape will change with depth of the image point, and degenerate into lines at the *tangential* and *saggital* focal points, with a circular PSF in between. At the tangential focal points, a wheel rim is sharply imaged, whereas the spokes would be clear at the saggital focus. The tangential and saggital focal surfaces are the locus of these focal points for all the points in the scene. Ideally these surfaces would be planar, but in general they will be curved, leading to curvature of field. The result is that it is impossible to image a plane such that points in the centre and periphery are simultaneously in focus.

The final Seidel aberration is distortion. Unlike the other aberrations, this assumes a sharp image of each point is formed on the image plane, such that there is no blurring. Instead, due to the position of aperture stops between the lens elements in a non-symmetric fashion, the chief rays of off-axis points may have a different magnification from those at the centre. The resulting image will be stretched inwards (*barrel distortion*) or outwards (*pincushion distortion*) at the corners. Although a SV PSF could describe this effect via a local shift, it is a purely geometrical effect and can thus be removed if necessary before considering deconvolution.

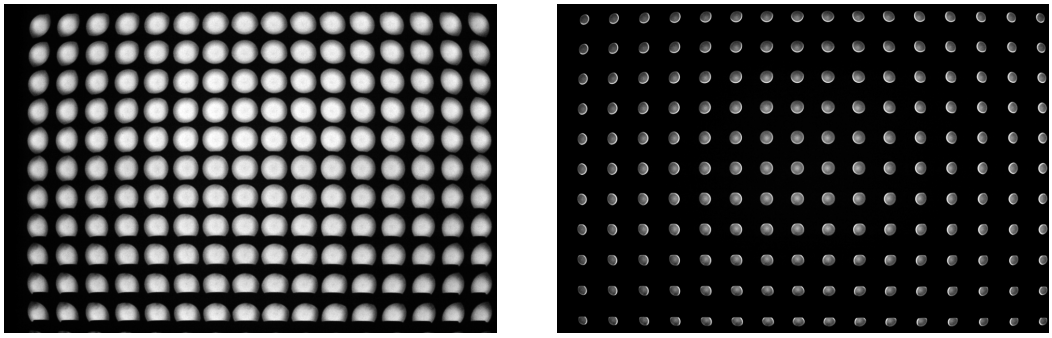


(a)



(b)

Figure A.5: *Lens aberrations. (a) shows an array of small white circles on a black background, photographed by a standard 50mm lens at f/8. Some of the circles have been enlarged for better visibility; it can be seen they are more or less uniformly circular and there are only minimal aberrations present; the 8 spikes that are just visible at the edges of each point in a star pattern are indicative of diffraction. (b) shows the diaphragm opened to give a wide aperture of f/1.4. The points towards the edges are now no longer circular but have trails pointing away from the centre of the image, which show evidence of coma.*



(a) (b)

Figure A.6: *Lens aberrations. The same target of small white circles as in fig. A.5 is used. (a) shows the target photographed at f/1.4 at a distance of 1m, with the focus set behind the target. In (b) shows the lens focussed in front of the target. The difference in shape and quality of the haloes in the PSFs at the centre may be attributed to spherical aberration. The "cat's eye" effect is also visible at the edge of the frame where some of the discs become clipped; those at the bottom are probably being clipped by the camera's mirror box, giving a straight edge.*

A.7 Compensating for aberrations

Aperture stops are used in lenses to limit the contribution of marginal rays to the image process. Use of smaller apertures reduces the effects of aberrations, tending towards the pinhole model again, but at the expense of light gathering ability. An alternative is for more complicated lens designs featuring many lens elements optimised to reduce the aberrations as much as possible, whilst still allowing wider apertures. These "fast" lenses tend to be heavier and more expensive than the simpler designs. Aberrations (other than distortion) that are not countered by the lens design will affect the sharpness of the resulting image, and can in general only be removed by SV restoration.

A.8 Chromatic abberations and vignetting

There are other effects that will influence the shape of the PSF. Optical vignetting is not an aberration as such, but an effect whereby different points of interest on the object have different aperture stops (i.e. different entrance pupils). In a real camera, the diaphragm is usually designed to be the aperture stop for all points of interest; however when fast lenses (low value of F) are being used, it is possible that any occlusions in the scene, or in the camera itself, may reduce the apparent aperture for points removed from the axis, and thus give a

different shaped PSF. Off-axis object points will in fact not see the whole of the entrance pupil that is visible to on axis points, but some other part of the lens barrel will block some of the marginal rays coming from these points. The result for an in-focus point is simply to let less light in, resulting in light fall-off towards the edges of the image, but for out-of-focus points, the PSF will have a clipped shape. This effect, visible in fig. A.6 is sometimes called the "cat's eye" effect [185].

Another aberration that is present, even with a first order analysis, is chromatic aberration. This is due to the variation of refractive index with wavelength of light, or *dispersion*, of the optical material. The result is that in a real scene where a distribution of wavelengths (colours) of light are present, different effective focal lengths and magnifications are obtained at the image plane. Using combinations of materials with different dispersion, to make different lens elements, in an *achromatic doublet*, can reduce this aberration. The magnification effects, or colour fringing (lateral chromatic aberration) may be reduced digitally in colour images simply by scaling the red, green and blue (RGB) colour channels independently.

All these effects should either be minimised prior to processing, or incorporated into the observation model. Chromatic aberration and distortion are fairly easily handled as corrective pre-processing steps, whereas the other distortions can be handled by a SV observation model where the PSF changes shape throughout the image field.

A.9 Physical optics models

Diffraction is another important effect which is not considered by the geometric optics models described so far. Studies have investigated theoretically [129, 209] and experimentally [198] the effects of geometric optics versus physical optics models, where diffraction effects are included. These effects become more critical when dealing with dimensions closer to the scale of the wavelength of light, but can have a pronounced effect on the resulting PSFs when smaller apertures are used.

Theoretical results suggest that as the degree of defocus increases, the two models become more similar. [198] have concluded that the differences are more pronounced at higher frequencies, and thus more important as higher spatial resolution sensors are used. Their (non-blind) restorations of real out-of-focus images taken under carefully controlled conditions did not show such significant differences; however with more advanced image models and restoration

procedures available today, these higher frequency details may be better restored, and thus the more accurate blur model may be useful. A further study in [192] suggests that at higher SNRs (above 30dB), the physical optics model is more useful.

These analyses have also been performed for circularly symmetric apertures; in practice camera apertures are often approximately polygonal based upon a number of overlapping blades. A simple way to consider the effects of diffraction from the aperture is that an additional convolutive component is introduced to the PSF: it is convolved with the Fourier transform of the aperture. Thus a circular aperture produces a Bessel function shaped PSF, even when the scene is perfectly focussed. Polygonal apertures will give star-like spikes as a component of the PSF. This places a limit on achievable resolution: the diffraction limited resolution is taken as the smallest line-pair that can be resolved (by separation of the first-zero crossings of the PSFs) when diffraction is the only limiting factor. In practice, there is an optimum aperture size, below which diffraction degrades the image, and above which lens aberrations and the chance for focus errors become more significant. The only way to resolve further details with a standard camera model is through the application of post-processing such as deconvolution.

A.10 Other degradations affecting digital images

The effects of the optical systems including lenses and apertures have been so far considered, but there are further stages necessary to build up the complete model of the photographic imaging process. Some of these effects are examined in more detail in books and reviews such as [79, 96, 129, 179]. The review article [129] was published before the widespread use of digital cameras, and therefore concentrates on the processes that occur in film-based cameras. These are still relevant, as one particular application of BIR is restoration of archival material (it has only been a few years since sales of digital cameras overtook those of film cameras), and will be considered here briefly before describing phenomena particular to digital cameras.

Fortunately the digital imaging process that has become common nowadays removes some of the non-linear effects caused by film, but introduces problems of its own.

A.10.1 Photographic film

In a photographic film camera, the image formed by light entering the lens exposes the film, causing an electro-chemical change in the silver-halide crystals in the film. During development, some of the crystals, or grains, are converted to silver, depending on the exposure received. This process is both non-linear and partially stochastic, resulting in *film grain* noise and a particular tone response curve. Colour films will also show some inter-channel effects.

In order to process photographs recorded on film digitally, they must be scanned somehow. Because of this most of the affects described below for digital imaging systems also need to be considered for restoration of film-based images. Scanners will of course introduce their own type of blur into the system due to their optics, and have different sampling, quantisation, and colour properties from a digital camera's sensor.

A.10.2 Pointwise nonlinearities

The non-linear response of film, considered at each point without any spatial interaction, may be described as a mapping between the incident light intensity (exposure, E) and the

Display devices also have a gamma response, so images must be compensated with the inverse response to display correctly. Digital image files are commonly stored in this gamma-compressed format, so this response should be removed before processing.

A.10.3 Camera shutter

[129] investigates the effect of the camera shutter on the final image. The shutter can be seen as a time-varying aperture which will affect the overall system response if it is not positioned exactly at the focal plane; however unless the shutter speed is very high, the effects will be negligible.

A.10.4 Sampling

Digital camera sensors, traditionally CCDs and more commonly now complementary metal oxide semiconductor (CMOS) devices, have to convert the analogue light field arriving at the focal plane into a spatially sampled electrical signal.

The effect of the image sensor may be modelled by an integration across the pixel region, followed by perfect sampling. [192] uses this model when investigating the accuracy of PSF modelling. For a square pixel, this effect may be approximated by a convolution with a square PSF. In real sensors, the integration region may need to be weighted by the actual response of the pixel. Each pixel typically has a microlens attached which attempts to focus light rays arriving from different angles onto the active area of the sensor. Furthermore, the edges of the pixel will typically not be light sensitive due to the other electronics needed to process the incoming signal, giving a *fill-factor* less than 100%.

A low-pass anti-aliasing filter is used in the majority of cameras, in front of the sensor. The response of this filter is unlikely to be perfect and will therefore introduce some additional blurring below the Nyquist frequencies into the system, which should be compensated for.

A.10.5 Quantisation

Many digital image file formats only store 8 bits quantised data. Even with RAW formats, the camera's analogue to digital converter (ADC) usually limits the precision to 10 to 14 bits precision. This also corrupts the signal — in a basic approximation we can also view this distortion as extra additive noise, which at low ISO speed (gain settings) will limit the BSNR.

A.10.6 Atmospheric Turbulence

For images taken outdoors over long distances, typically landscape images, or images looking towards the sky, random motions of particles in the atmosphere changing refractive index add an extra blur to the observed image, which is often modelled being as Gaussian when averaged over time. This type of blur places limits on astronomical observations from the earth's surface.

A.10.7 Colour filter array

The Bayer pattern that most digital cameras feature on their sensor provides a way to capture colour images on a single sensor, by filtering light into the three primary colour channels at different spatial locations. That is, effectively a reduced sample rate is used to capture each channel, with twice the resolution used for Green as the other two channels. This process means interpolation of the data is usually necessary to reconstruct a colour image; this interpolation introduces extra blurring into the final image.

A.10.8 Compression

Observe that whilst many of these effects, such as photonic noise, are unavoidable, others may be reduced through different design of the camera processing system. For example, obtaining RAW data from the ADC is preferable than using a compressed, over-quantised version if we wish to process the image further. Otherwise, these effects should be considered as extra parts of the observation model or extra tasks for restoration.

A.11 Noise models

Noise covers any other stochastic perturbation of the signal away from the deterministic part of the model. It attempts to account for outside interference and other physical effects causing measurement uncertainty, for example electronic, photoelectric, film grain, and quantisation noise.

In digital imaging, thermal noise in the photodetector and subsequent processing circuitry is one of the main sources of noise. This is well modelled by an additive Gaussian process, $w(\mathbf{x})$, usually assumed to be WGN. However there may be cross-talk between pixels, or other filtering effects in the analogue circuitry and ADC which cause the noise to be correlated. In colour imaging, the noise may also be separated into luminance and chrominance noise, describing monochromatic noise and noise that is correlated across the channels.

Digital sensors may also suffer from fixed pattern noise, where certain pixels may have additional gains or offsets, especially under long exposures. This type of noise is easily dealt with by subtracting a reference "dark frame" where no light is present.

At very low signal levels, the quantum nature of light must be taken into account. This means that the signal will be quantised based on the number of incident photons arriving at the detector, and hence the actual received signal, $g_r(\mathbf{x})$, is modelled by a Poisson distribution

$$p\left(g_r(\mathbf{x}) \,|\, g_0(\mathbf{x})\right) = \frac{g_0(\mathbf{x})^{g_r(\mathbf{x})} \exp\left(-g_0(\mathbf{x})\right)}{g_r(\mathbf{x})!}, \tag{A.13}$$

given the time-averaged incident illumination (plus background radiation/dark current) level, g_0. In practice, this is of more concern in applications such as medical imaging where extremely high gains are used, for instance with photomultiplier tubes [179]; however at the increasingly high ISO speeds being made available in today's digital cameras, the Poisson model

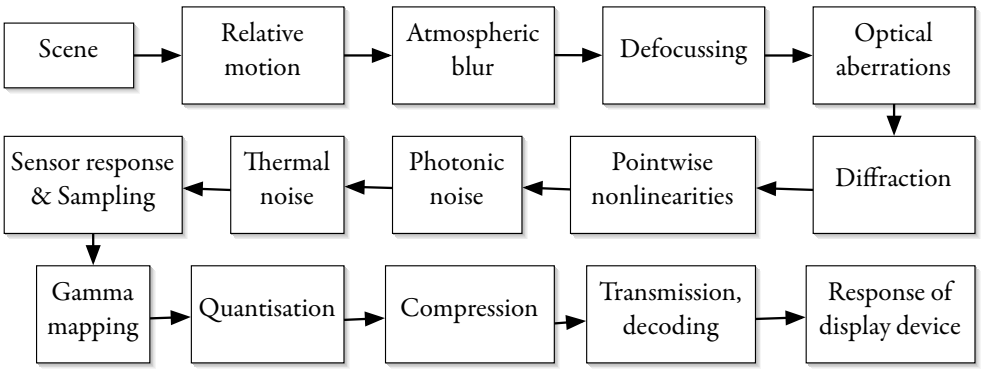


Figure A.7: *Block diagram showing degradation processes in a digital imaging system*

may also be appropriate.

At higher signal levels, the Poisson noise is well approximated by a Gaussian distribution, with mean and variance both equal to the signal level. It is also possible to arrive at the Gaussian noise model by invoking the central limit theorem: there are many different noise sources which when added together will tend towards a Gaussian distributed variable.

A.12 Complete imaging process

Putting together the various effects that appeared in the previous sections, we may produce a system model for the imaging process, as shown in fig. A.7.

Having considered all these effects, we choose to form a simplified view of the problem and linearise some of the distorting effects. We will assume that other effects can either be absorbed into the linear model, treated as noise, removed in pre-processing, or are negligible for the applications of interest. The model will directly represent the defocussing and other blurring effects, including non-geometric aberrations, assuming monochromatic light, a linear sensor response, and an uncompressed, unquantised, ideally sampled image with only additive noise.

A.13 Conclusions

In this appendix, an overview of the photographic imaging process was described, and factors that may influence the PSF model were discussed.

Appendix B
MRF and other image models

In this appendix, we give some further background details on Markov random field (MRF) models and their use in image processing, as well as the different forms they can appear in, before summarising a selection of other image models. Note that many of the image models that have been used in BID, including the ones we propose in the thesis can be considered as MRFs.

B.1 MRF approaches in image processing

The MRF may be considered a generalisation of the Markov chain (MC) to 2D. They arose in the study of lattice systems in statistical physics in the 1950's and 60's, and were first applied to image processing in the 1970's and 80's. The discrete MRF modelling spatial interactions in 2D was presented in, for example, [229], and more generally in [19], while the first applications to image restoration were in [56, 136, 230] and the influential paper of Geman and Geman [76]. This, as well as later work such as [99], makes use of compound or hierarchical fields to better model the images. Molina *et al.*have recently applied the CGM type models to the single [148–150] and Multichannel [152] restoration problem using methods such as SA and ICM.

Once the models have been defined, a variety of approaches such as recursive filtering or Bayesian methods such as MAP estimates of the conditional image field are sought. Often these lead to very large computations so approximate estimation methods are required, such as the ICM method [20], or Mean Field Theory [238], similar to the VB methods we use in this thesis. [76] combines a number of key concepts, including numerical Bayesian methods (Gibbs Sampling) and an iterative procedure for convergence (SA or Stochastic Relaxation) that were, again, first used in statistical physics. It should be noted that there is a strong relation between MRF models and other types of image priors (such as those mentioned for TV like methods); AR models with Gaussian excitation are a special case of MRFs (see [56]).

MRFs have also been used extensively for image segmentation (which is somewhat related to image modelling).

B.2 Mathematical MRF formulation

A detailed description of definitions and applications of various MRF models is given in [226] and tutorial notes on the topic are presented in [33], some aspects of which are summarised here. As already mentioned, these models developed from ideas in statistical mechanics and were introduced in important papers including [19, 56, 76, 229, 230], however their widespread use is only since the work of Geman and Geman [76]. These models are sometimes also called conditional Markov (CM) [56] or CAR models [154, 155].

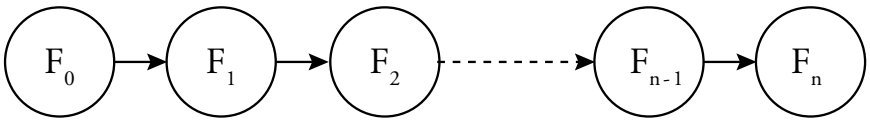


Figure B.1: *A Markov Chain*

We may consider MRFs a generalisation to 2D of Markov chains (MCs). The upper-case notation for random variables and lower-case for particular states of these variables will be used in this section, and bold to indicate a vector or set of variables. A MC (fig. B.1) is a sequence of random variables $\{F_0, F_1, \cdots, F_n\}$ defined with conditional probabilities according to the *Markov property*, that the state of a variable F_{i+1}, given any subset ε of its predecessors, is only influenced by its immediate predecessor F_i:

$$p\left(F_n = f_n \,\middle|\, \left\{\varepsilon \subset \{F_i = f_i : i < n\}\right\}\right) = p(F_n = f_n \mid F_{n-1} = f_{n-1}) \tag{B.1}$$

or equivalently its joint PDF representation may be found, due to the causal dependencies, simply by multiplying the conditionals:

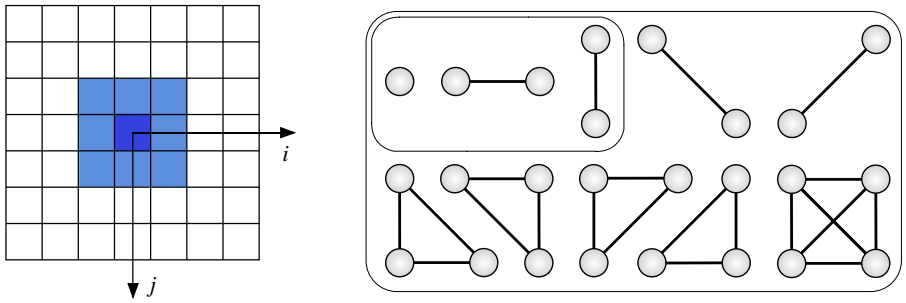
$$p\left(\mathbf{F} = \mathbf{f}\right) = p(F_0 = f_0) \prod_{i=1}^{n} p(F_i = f_i | F_{i-1} = f_{i-1}), \tag{B.2}$$

where $\mathbf{F} = \{F_i : 1 \leqslant i \leqslant n\}$. Note that higher order dependencies may be expressed with a Markovian representation by extending the state vector to include the p previous states which influence the current state, and the model is then termed p^{th}-order Markov or Markov-p. It

is common to use Markov chains for time series representations, where the index i becomes a time variable t.

Unfortunately in two dimensions, definition of the MRF becomes slightly more difficult due to lack of natural ordering and the ill-defined notion of causality. It is first necessary to introduce some preliminary concepts. The random field $\mathbf{F} \triangleq \{F_s : s \in \Omega\}, \mathbf{F} \in \mathbb{A}_F$, is defined on a finite 2D lattice structure with *sites* at pixel locations $s \triangleq (i, j)$, and these are indexed by the set $\Omega \triangleq \{s \in \mathbb{Z}^2 : 1 \leqslant i \leqslant m, 1 \leqslant j \leqslant n\}$. The state space of a particular site F_s is discrete and comes from alphabet $\mathbb{A}_{F_s}$, i.e. the pixel has a finite set of gray levels ($\mathbb{A}_{F_s}$ is assumed common to all sites). Similarly, all the possible images or configurations for the random field are denoted $\mathbb{A}_F$.

We may then define a conditional *neighbourhood system*, $\mathcal{N} \triangleq \{\mathcal{N}_s : s \in \Omega\}$, where each site s has a *symmetric neighbourhood* $\mathcal{N}_s \triangleq \{r \in \Omega : s \in \mathcal{N}_r \Leftrightarrow r \in \mathcal{N}_s; s \notin \mathcal{N}_s\}$ which is homogeneous in shape except at image borders. It is usual to use circular neighbourhoods, where a d^{th} order neighbourhood is given by $\mathcal{N}_s^d = \{r \in \Omega : 0 < |s - r| \leqslant d\}$. A second order neighbourhood is shown in fig. B.2(a). The neighbours of s represent the pixels which



(a) 2nd order Neighbourhood (lighter coloured) of central pixel

(b) 2^{nd} order cliques (1^{st} order shown in inner group)

Figure B.2: *Neighbours and cliques for 2nd order Neighbourhood*

influence s and vice versa; hence they may be used in the conditional probability definition of the MRF, that the probability of variable F_s taking gray level f_s, given any other subset ε of the field $\mathbf{F}$, only depends on its neighbours:

$$p\left(F_s = f_s \,\middle|\, \{\varepsilon \subset \{F_r = f_r : r \neq s\}\}\right) = p(F_s = f_s \mid F_r = f_r : r \in \mathcal{N}_s) \tag{B.3a}$$

$$\text{and} \quad p(\mathbf{F} = \mathbf{f}) > 0, \quad \forall \mathbf{f} \in \mathbb{A}_F \tag{B.3b}$$

where for a complete definition and proof of the theorem in (B.5) it is usual to add the *positivity condition* (B.3b).

Furthermore, if the notions of *cliques* and *potentials* are introduced, we may define a *Gibbs distribution* w.r.t $\mathbf{F}$ given $(\Omega, \mathcal{N})$. Cliques allow the interaction between sites to be specified; they are either subsets of the neighbourhoods, or single sites. Every pair of sites in a clique are mutual neighbours, i.e. a clique is any set $c \subseteq \Omega$, where $c = \{r \in \Omega : (s \in c, r \neq s) \Rightarrow r \in \mathcal{N}_s\}$. Each pixel may be a member of several cliques. The set of all cliques in the neighbourhood system is $\mathcal{C} \triangleq \{c \subseteq \Omega \mid \mathcal{N}\}$, and all cliques containing site s is $\mathcal{C}_s \triangleq \{c \subseteq \mathcal{C} : s \in c\}$. The cliques for the second order neighbourhood are shown in fig. B.2(b). A potential function $V_c(\mathbf{F})$ is *any* function defined on the lattice Ω, such that it only depends on pixel values f_r at sites r that are members of the clique c, i.e. $V_c(\mathbf{f}) = V_c(\{f_r : r \in c\})$. Then given the set of potentials, the Gibbs distribution is a joint PDF defined as:

$$p(\mathbf{f}) = \frac{1}{Z} \exp\left[-\frac{1}{T}\sum_{c \in \mathcal{C}} V_c(\mathbf{f})\right] = \frac{1}{Z} \exp\left[-\frac{1}{T} U(\mathbf{f})\right] \tag{B.4}$$

where $Z \triangleq \sum_{\mathbf{f} \in \mathbb{A}_F} \exp[-U(\mathbf{f})]$ is the normalising constant (also known as a *partition function*) and $U(\mathbf{f})$ is the sum of all potentials, called the *energy function*. The *temperature* parameter T is sometimes introduced (normally $T = 1$) in analogy to statistical physics and alters the amount of "peaking" in the distribution.

Finally, it is possible to complete the analogy to the 1D MC case, and provide a form ready for use, by specifying the link between the conditional definition of the MRF (B.3a) and the joint PDF of the Gibbs distribution (B.4). It is not immediately obvious how to procede; with non-causal dependencies, factoring a joint PDF into conditionals is difficult. Fortunately, this is possible due to the fundamentally important *Hammersley-Clifford Theorem*:

$$\mathbf{F} \text{ is an MRF} \iff p(\mathbf{f}) \text{ is a Gibbs distribution.} \tag{B.5}$$

The proof is given in [226], and is originally due to Besag [19]. With this theorem it is possible to link local and global properties of the MRF, and specify the model in a natural form, i.e. with local conditional probabilities, then work with the PDF of the whole image (or to find conditionals as may be required by the Gibbs sampler from a joint pdf). The conditionals, or

local characteristics may be found as:

$$p\left(F_s = f_s \,|\, \{F_r = f_r : \forall r \neq s\}\right) = \frac{p(\mathbf{f})}{p(\mathbf{f} \setminus f_s)} = \frac{p(\mathbf{f})}{\sum\limits_{f_s \in \mathbb{A}_{F_s}} p(\mathbf{f})} \tag{B.6}$$

$$= \frac{\exp\left[-\sum_{c \in \mathcal{C}_s} V_c(\mathbf{f})\right]}{\sum\limits_{f_s \in \mathbb{A}_{F_s}} \exp\left[-\sum_{c \in \mathcal{C}_s} V_c(\mathbf{f})\right]} \tag{B.7}$$

where the definition of the Gibbs distribution has been used assuming $T = 1$, and if the state space were not discrete, the summation in the marginalisation of f_s would be replaced by an integral. The denominator may again be regarded as a normalising constant.

A key benefit of the MRF definition is the ability to specify almost arbitrarily complex models through small numbers of parameters and minor extensions to the model. A problem is how to specify the conditionals, or equivalently how to choose the potentials. Any function may be used, and different types of interaction can be specified by weighting or omitting certain types of cliques. It is noted in [19] that a variety of standard forms, called *auto-models* may be obtained using PDFs from the exponential family and only pairwise interactions (cliques with two sites); then it is possible to factorise the potentials into a product of two identical terms, which are linear in the pixel values. A similar idea is shown in [193, 194] for a particular standard case, the GMRF. Then the Gibbs distribution becomes a multivariate Gaussian:

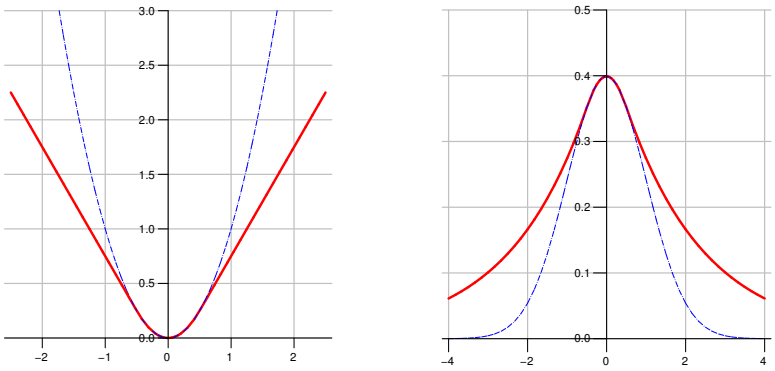
$$p(\mathbf{f}) = \frac{1}{Z} \exp\left[-\mathbf{f}^T \mathbf{B} \mathbf{f}\right] = \frac{1}{Z} \exp\left[-\sum_{c \in \mathcal{C}} \mathbf{f}^T \mathbf{B}_c \mathbf{f}\right] \tag{B.8}$$

where the inverse covariance matrix $\mathbf{B}$ is defined by appropriately chosen clique weights; entries[1] $[\mathbf{B}_c]_{s,r}$ are only non-zero when pixels s and r are neighbours. This decomposition is possible for arbitrary sized cliques, and for some sizes a factorisation $\mathbf{B} = \mathbf{D}^T\mathbf{D}$, or equivalently $\mathbf{B}_c = \mathbf{d}_c \mathbf{d}_c^T$ is possible where $\mathbf{D} = [\mathbf{d}_{c_1} \mathbf{d}_{c_2} \cdots \mathbf{d}_{c_{|\mathcal{C}|}}]^T$, and $\mathbf{d}_c$ is only nonzero at entries $[\mathbf{d}_c]_s$, $s \in c$. Typically first or second order finite difference approximations are used for the operators $\mathbf{d}_c$. However with the assumptions made for the auto-models, the factorisation using cliques of size two is always possible, which turns out to be expressable in the form $\mathbf{f}^T\mathbf{B}\mathbf{f} = \sum_{c \in \mathcal{C}} V_c(\mathbf{f}) = \sum_{\substack{c \in \mathcal{C} \\ c = \{s,r\}}} \alpha_s f_s^2 + \beta_{sr}(f_s - f_r)^2$ (the α_s are usually zero).

[193] then extends this idea to GGMRFs with arbitrary non-quadratic potentials of a similar functional form: $V_c(\mathbf{f}) = \rho(\mathbf{d}_c^T \mathbf{f})$, where ρ is some (usually convex) function, such as the

[1] A slight abuse of notation: this really means lexicographically ordered entries at row $N \cdot i_s + j_s$ and column $N \cdot i_r + j_r$ of $\mathbf{B}_c$.

Huber function shown in fig. B.3, which leads to a Huber MRF (HMRF). The idea is to use



(a) Huber and quadratic functions

(b) Huber and Gaussian PDFs

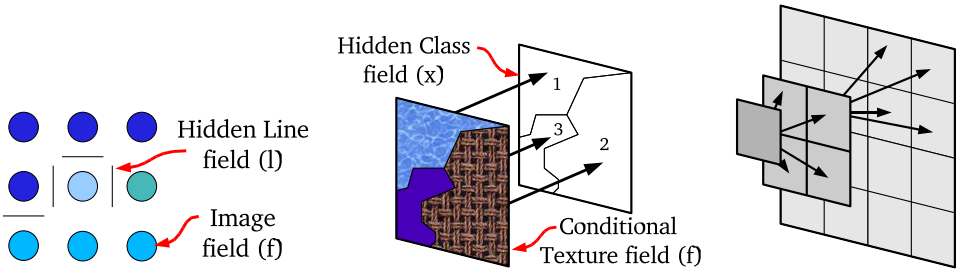
Figure B.3: *The Huber function (solid), which is quadratic below a threshold (0.5 in this case), and linear above. Quadratic* x^2 *shown (dashed) for comparison.*

a heavy-tailed distribution to penalize outliers (such as jumps at edges) less severely than the GMRF which tends to oversmooth when used as a prior.

Various authors have suggested using compound or hierarchical MRF models, in which there are multiple levels of random fields, some of which may be hidden in analogy to 1D Hidden Markov Models (HMMs), and represent unobserved elements in the images such as the class of a pixel or the location of an edge. These models appear to hold great promise for image restoration and segmentation tasks. For example Geman and Geman [76] use a hidden *line field* to model the edges, where the correlation between pixels (i.e. the potentials) is determined by presence or absence of lines in different directions as in fig. B.4(a). The compound Gauss-Markov (CGM) is an extension of this model from discrete to continuous gray levels and is described, along with the doubly stochastic Gaussian (DSG) model in work by Jeng and Woods [98, 99]. The DSG is a 'switched' process such that there are several underlying model parameters which again depend on the value of the underlying line field, however in this case both the image and line models are causal such that the line process and the combined model are both MRFs.

This DSG model seems to resemble methods in spatially adaptive restoration where the amount of regularisation is dependant on the local activity. However ideas from work on textured image segmentation may be more general; rather than just modelling the activity near and away from edges, they have a different possible upper level model in each region, depending on the

class of the lower level region (see fig. B.4(b)). Bouman and Shapiro [35] have referred to this idea as a doubly stochastic random field, which is used the basis of more advanced *multiscale* approaches, where there image is also represented as an 3D MRF in scale (fig. B.4(c)). This is often a causal representation from less to more detailed levels, and allows for larger level spatial interactions to be specified using fewer parameters than a single level model. Further ideas along these lines are discussed in [226]. Application of these models to image restoration may yield improved results.



(a) The CGM Line Field (b) Doubly stochastic model (c) Multi-scale MRF

Figure B.4: *Multi-level or Hierarchical random fields*

B.3 MRF discussion

It is clear there is a distinct similarity between the models mentioned and also the many forms of prior terms used in the literature. Chellappa *et al.*[56] attempt to compare the SAR and CM (i.e. GMRF) models, and assert that neither is a subset of the other. However this turns out to be not entirely true for the case of GMRFs and AR models with Gaussian excitation: they are equivalent in form; only the covariance is interpreted in a different way.

[56] shows that the GMRF model may be written as a recursion, similar to the AR model, except that the excitation noise $\nu(s)$ is not white, but coloured. Specifically, if the model parameters are b_r (akin to the $a(i-k;j-l)$ in the AR model), the noise has correlation

$$E[\nu(s)\nu(r)] = \begin{cases} \sigma_\nu^2, & s = r \\ -b_{s-r}\sigma_\nu^2, & r \in \mathcal{N}_s \\ 0, & \text{otherwise} \end{cases} \tag{B.9}$$

The recursion may be written in the matrix-vector form $\mathbf{B}\mathbf{f} = \boldsymbol{\nu}$, also we may write $\mathbf{B} = \mathbf{I} - \mathbf{B}'$ as in the AR model. It is then shown that necessary conditions for stability in the bounded input bounded output (BIBO) sense are positivity of the eigenvalues of $\mathbf{B}$, i.e. $\mathbf{B}$ positive definite ([19] also mentions this requirement). Furthermore, $\mathbf{B}$ is symmetric from the definition of the neighbour sets. This implies a factorisation is always possible via the Cholesky decomposition[2], $\mathbf{B} = \mathbf{L}^{\mathsf{T}}\mathbf{L}$. As $\mathbf{L}$ is a triangular matrix, it may be used to represent a causal AR process, $\mathbf{L} = (\mathbf{I} - \boldsymbol{A})$, although generally the coefficients will be non-stationary and so will the excitation variance as the main diagonal may not contain ones. Hence the non-causal GMRF representation is likely to be more efficient.

It may be seen why the noise is non-white for the MRF by considering the covariance matrices or PDFs. For the AR model we have $\mathrm{p}(\mathbf{f}) \propto \exp\left[-\frac{1}{2}\mathbf{f}^{\mathsf{T}}(\mathbf{I} - \boldsymbol{A})^{\mathsf{T}}(\mathbf{I} - \boldsymbol{A})\mathbf{f}\right]$ and $\mathrm{p}(\boldsymbol{\nu}) = \exp\left[-\frac{1}{2}\boldsymbol{\nu}^{\mathsf{T}}\boldsymbol{\nu}\right]$ via probability transformation $\boldsymbol{\nu} = (\mathbf{I} - \boldsymbol{A})\mathbf{f}$. For the GMRF we have $\mathrm{p}(\mathbf{f}) \propto \exp\left[-\frac{1}{2}\mathbf{f}^{\mathsf{T}}(\mathbf{I} - \mathbf{B}')\mathbf{f}\right]$, and hence $\mathrm{p}(\boldsymbol{\nu}) = \exp\left[-\frac{1}{2}\boldsymbol{\nu}^{\mathsf{T}}(\mathbf{I} - \mathbf{B}')^{-1}\boldsymbol{\nu}\right]$.

In summary, as [56] mentions, it is simple to convert the Gaussian SAR model to an MRF, simply by finding the product $\mathbf{B} = (\mathbf{I}-\boldsymbol{A})^{\mathsf{T}}(\mathbf{I}-\boldsymbol{A})$, and the reverse process of factorisation into an AR form *is* more difficult but always possible. In the non-Gaussian case an MRF model with the same second-order properties as the SAR model is available, but possibly not an exact match. Again for a non-Gaussian MRF model there may not be an equivalent *simple* stationary SAR model of the kind they describe (in this sense general MRF models are a superset of simple SAR models with iid WGN), but there is in fact always a (possibly nonstationary) causal AR factorisation. This is demonstrated by Brook [36]: the AR representation is essentially a *joint* one, whereas the MRF is essentially a *conditional* representation; though this may be converted into a joint PDF via the Hammersly-Clifford theorem for any Gibbsian PDF. In fact [36] shows that the conditional representation is *necessarily* a simpler type of model and a special case of the more complex joint specification: the conditional form is *always convertable into a 1D causal Markov Chain* (though analytically finding its transition probabilities may be difficult). The conditional form imposes certain stringent consistency constraints; however these may be useful in practice in specifying useable modells as is argued by Besag [19] (i.e. the full joint description is too general). Thus it is concluded that the MRF models are in fact useful, and though they may be specified through either joint or conditional densities, they

[2]This may not be the most efficient factorisation, and could involve a semi-infinite support (i.e. the size of the image), whereas in certain cases other simple (possibly non-causal) factorisations are possible. A example of the simple case is given in fig. 2 of [56]. Also, the factorisation $\mathbf{B} = \mathbf{D}^{\mathsf{T}}\mathbf{D}$ already mentioned is always possible in the Gaussian case, for pairwise cliques, although the resulting matrix $\mathbf{D}$ may be non-square.

are of the simple kind described by Brook which constrain the form of their joint probability (i.e. a Gibbs distribution).

An advantage of using MRF descriptions is that they are closely linked to various numerical estimation and simulation methods first derived for large scale processes from statistical physics, such as MCMC methods. Methods such as the Metropolis algorithm or Gibbs sampling and SA may be applied fairly straightforwardly. There is a large body of work available this area in providing rigorous solutions to difficult estimation problems.

Finally, it is noted that the prior terms used in standard regularisation theory, e.g. discrete approximations to the Laplacian operator, and various TV and anisotropic functionals may be considered special cases of AR or MRF models. The Laplacian is clearly seen to be an AR model with coefficients of 0.25 on the four adjacent pixels. The κ function from TV regularisation is clearly likened to the ρ function in the potentials of generalised GMRFs; the Huber function and the method of You and Kaveh [234] appear to have the same motivation in their choice of function. Molina *et al.*[155] describe the difference between the *first difference* constraints used in the CAR (i.e. GMRF model) and *second difference* constraints imposed by the same operator using an SAR form. This is somewhat analogous to the difference between the standard use of the Laplacian and the terms used in TV. One other point of note: choice of actual image model parameters has been something of a difficult issue in most previous regularisation methods, which often fall back to ad-hoc methods and it is expected that methods that jointly estimate model parameters with the image and blur should provide better results.

B.4 Anisotropic Diffusion and Total Variation Type Models

Non-quadratic image priors have been investigated using *variational integrals* in the *anisotropic diffusion* [236] or *total variation (TV)* [50] regularization frameworks, with the aim of preserving edges by not over-penalizing discontinuities, i.e. outliers in the image gradient distribution, see [49, 85] for a unifying view of the probabilistic and variational approaches. The main difference to the GGMRFs models mentioned in§4.4.3 is that these usually begin with a formulation in the continuous image domain resulting in partial differential equations (PDEs) that must be solved. However, eventual discretization is eventually necessary, and hence the constraints may be reformulated as non-Gaussian priors, or Gaussians with a non-stationary covariance matrix [206]. Alternatively, other methods propose formulating the TV norm di-

rectly in the discrete domain [22, 52].

The generalized regularization approach using anisotropic diffusion has been proposed by You and Kaveh [236]. In this formulation, convex functions $\kappa(\cdot)$ and $\upsilon(\cdot)$ of the image gradient $|\nabla f(x)|$ and the PSF gradient $|\nabla h(x)|$ respectively are used in defining regularization functionals:

$$\mathcal{E}(f) = \int_{S_f} \kappa\left(|\nabla f(\mathbf{x})|\right)\, d\mathbf{x} \tag{B.10a}$$

$$\mathcal{E}(h) = \int_{S_h} \upsilon\left(|\nabla h(\mathbf{s})|\right)\, d\mathbf{s}. \tag{B.10b}$$

This is in analogy with standard regularization procedures. However as the functionals are continuous, variational calculus is used to perform the differentiation needed for minimization [237]. This results in a PDE which must be solved for each variable. Consider for instance minimization of Equation (B.10a) for f; the solution must satisfy

$$\bigtriangledown_f\, \mathcal{E}(f) = \bigtriangledown \cdot \left(\frac{\kappa'(|\bigtriangledown f|)}{|\bigtriangledown f|} \bigtriangledown f \right) = 0, \tag{B.11}$$

with appropriate boundary conditions. One method of solution is imposing an artificial time evolution variable t, and using a steepest descent method, i.e., for f,

$$\frac{\partial \hat{f}}{\partial t} = -\bigtriangledown_f\, \mathcal{E}(\hat{f}). \tag{B.12}$$

This may be interpreted as representing a physical anisotropic diffusion process [224, 236, 237]. That is, as the time variable t progresses, directional smoothing occurs depending on the local image gradient. The strength and type of smoothing depends on the *diffusion coefficient* or *flux variable*, c, which is related to the potential function by

$$c(|\bigtriangledown f|) = \frac{\kappa'(|\bigtriangledown f|)}{|\bigtriangledown f|} \tag{B.13}$$

We may consider $c(|\bigtriangledown f|)$ to be the amount of smoothing perpendicular to the edges. Appropriate choice of c or equivalently κ can result in spatially-adaptive edge preserving restoration.

Consider two cases of the potential function κ and related diffusion coefficient c. In the first case, $\kappa(x) = \frac{1}{2}x^2$ and hence $c(|\bigtriangledown f|) = 1$, and $\bigtriangledown_f \mathcal{E}(f) = \bigtriangledown^2(f)$, i.e., a Laplacian operator [234]. This corresponds to standard spatially-invariant isotropic regularization, or a CAR model with

the discrete Laplacian when discretized.

Another choice proposed for the BID problem by Chan and Wong [50] is given by the total variation (TV) norm. In this case, $\kappa(x) = x$ and hence $c(|\bigtriangledown f|) = \frac{1}{|\bigtriangledown f|}$. The result is that smoothing in the direction orthogonal to the edges is completely suppressed and is only applied parallel to the edge directions. This is demonstrated in [236] by decomposing the anisotropic diffusion equation, Equation (B.12) into components parallel and perpendicular to the edges. A very efficient way to solve the resulting optimisation problem is shown in [50] in this particular case.

These two choices lead us to consider the following discrete prior image models:

$$p(\mathbf{f}) \propto \exp\left[-\alpha_{im} \sum_i ((\Delta_i^h \mathbf{f})_i^2 + (\Delta_i^v \mathbf{f})_i^2)\right] \tag{B.14}$$

for the Laplacian; and

$$p(\mathbf{f}) \propto \exp\left[-\alpha_{im} \sum_i \sqrt{(\Delta_i^h \mathbf{f})_i^2 + (\Delta_i^v \mathbf{f})_i^2)}\right] \tag{B.15}$$

for the TV norm, where Δ_i^h and Δ_i^v are linear operators corresponding to horizontal and vertical first order differences, at pixel i, respectively.

A combination of the two choices for c is considered in [236], resulting in a spatially-adaptive diffusion coefficient. The smoothing strength is increased using the Laplacian in areas with low gradient magnitude, and decreased using the TV norm in areas where large intensity transitions occur in order to preserve edges while still removing noise. This is analogous to the use of the Huber function in Section 4.4.3. Many other diffusion coefficients are proposed in the literature, including very complex structural operators (see [224] for a review).

Šroubek and Flusser [206] use a similar scheme to those already mentioned, but (using the half-quadratic approach [45, 74]) demonstrate how the anisotropic diffusion model may be written in the form of Equation (4.31) by discretization of the functional in Equation (B.10a):

$$p(\mathbf{f}, c(\mathbf{f})) = \frac{1}{Z_f} \exp\left[-\frac{1}{2}\mathbf{f}^T \mathbf{B}(c)\,\mathbf{f}\right] \tag{B.16}$$

They equate the diffusion, or flux variable, to the hidden line process often used in CGMRFs, that is it represents the edge strength between two pixels in the image. Therefore it is possible

to build a spatially-varying weights matrix **B** from the local image gradients. Note that as the flux variable is a function of **f**, so is the covariance, so this is not strictly a Gaussian distribution unless **B** is assumed fixed. In practice, using an iterative scheme, **B** may be updated at each iteration.

A similar motivation was used in [235] to obtain a spatially-varying weights matrix based on the local image variance, as was previously suggested in [109, 121]. The difference here is that the regularization is isotropic; better performance can be expected with the anisotropic schemes.

B.5 Other Image Models

In this section, several other image models, or variations on those already discussed, will be summarised; they have not been used specifically for BID but have been proposed for related problems such as denoising, classical restoration or inpainting.

B.5.1 Recursive Filtering

Recursive filtering is another possibility that enables some degree of spatially-varying restoration as well as reduced storage requirements since local processing is used. The Wiener filter may be reformulated recursively as a 2D discrete Kalman Filter [227, 228]. An AR model for the image, representing prior knowledge of the image statistics is built, and in conjunction with the expression for the blurred observation, the problem is reformulated in state-space form. The Kalman filter may be extended to a non-linear and non-Gaussian formulation via the Particle Filter. However recursive approaches require a causal ordering of the data, which is not the most natural ordering for images.

B.5.2 Richardson-Lucy

The method developed by Richardson and Lucy [188] has been used successfully in astronomical imaging and microscopy; it is essentially an application of the EM algorithm [60] to the ML estimation of an image under the assumption of poisson, or multiplicative[3] noise [218] (which models the photon noise in low-light imaging, but may be applied in other cases too).

[3]The poisson noise may be approximated as a non-stationary Gaussian noise with variance equal to the mean, i.e. a unit Gaussian multiplied with the signal

While the ML problem is still ill-posed, again early termination of EM iterations may provide a form of regularisation.

B.5.3 Maximum Entropy

Another class of algorithms that are particularly popular in astronomy [151], maximum entropy methods (MEM), first developed by Frieden [68], may be compared to the methods already mentioned either as a constrained optimisation or a Bayesian restoration, except where a measure of *entropy* is used as the constraint or image prior [215]. The image entropy, defined analogously to the probabilistic measure as $\mathcal{E}(\mathbf{f}) = -\sum_{i,j} f(i,j) \ln(f(i,j))$ (which is maximised in the presence of no other constraints by a uniform image) again favours smooth images. By definition, the principle also has the advantage of favouring non-negative images.

B.5.4 Wavelet based methods

Formulating the restoration problem using a multiresolution or subband approach, dealing with features at different scales provides an extra way to model properties of images. One such approach to image restoration was given in [11], which applies multichannel restoration (see §C.9) to the wavelet subbands. Other more recent approaches, e.g. [83, 177] have considered using empirical modelling of the statistics of subbands in an overcomplete transformed domain, which provides a strong model for correlations in the signal, and good results in denoising applications.

B.5.5 Dictionary learning methods

Rather than using a known transform as in wavelet methods, where it is hoped that the signal will be sparse (that the energy will mostly exist in only a few coefficients), other methods have proposed learning overcomplete dictionaries of basis functions that well represent natural images. Methods such as [64, 134] seek both a sparse representation and an optimal dictionary to represent the signal, which can give good results in applications such as denoising. In [190], a MRF type prior is learned from a database of images, using training of both filters and weights on the potential functions. This type of method seems particularly powerful, and as a MRF could be applied in most Bayesian restoration contexts.

Appendix C
Review of Non-Bayesian BID approaches

In this appendix, other BID models, which have appeared in the literature and cannot be obtained from the Bayesian formulation described in previously, are briefly reviewed. Also included are discussions of multichannel and SV deconvolution.

C.1 Spectral and Cepstral Zero Methods

These algorithms fall into the *a priori* class of approaches described in §5.1, according to which the PSF of the blurring system is estimated separately from the image. They were the first examples of BD to be developed for images (see Stockham *et al.* [208] and Cannon [42]).

These blur identification algorithms are well suited to the problem when the frequency response of the blurring system has a known parametric form that is completely characterized by its frequency domain zeros: this is the case for linear motion blur and circular defocusing blur (Section 4.3.1). Let us rewrite Equation (2.6) in the frequency domain, while ignoring the noise term, i.e.,

$$G(\omega) = F(\omega) H(\omega). \tag{C.1}$$

Due to the multiplication of the spectra, the zeros of $G(\omega)$ are the zeros of $F(\omega)$ and $H(\omega)$ combined. Therefore the problem reduces to identifying which of the zeros of $G(\omega)$ belong to $H(\omega)$. The use of the parametric model makes this possible: the FT of the linear and circular blurs are Sinc and Bessel functions that have periodic patterns of zeros in their spectra. The spacing between the zeros depends on the parameter L in Equation (4.27) or r in Equation (4.28). Hence if this pattern can be detected in $G(\omega)$, the parameters can be identified.

In practice this type of identification often fails due to the presence of noise masking the periodic patterns. The *homomorphic* [208] or *Cepstral* [42] methods attempt to exploit the effect of the nonstationary image and stationary blur to alleviate this difficulty. The homomorphic

procedure begins by partitioning the image into blocks f_i, of size larger than the PSF. Each blurred block g_i is equal to the blur convolved with the unblurred block:

$$g_i(\mathbf{x}) = f_i(\mathbf{x}) * h_i(\mathbf{x}) + n_i(\mathbf{x}). \tag{C.2}$$

This expression is true, apart from at the block boundaries, due to contamination from neighbouring blocks (in practice the blocks are windowed to reduce these edge effects). The log operator may then be applied to the FT of the blurred image, which has the effect of converting the original convolution to an addition. The average of these blocks may then be calculated (assuming N blocks):

$$\frac{1}{N}\sum_{i=1}^{N} \log(G_i(\boldsymbol{\omega})) \approx \frac{1}{N}\sum_{i=1}^{N} \log(F_i(\boldsymbol{\omega})) + \log(H_i(\boldsymbol{\omega})). \tag{C.3}$$

This summation now consists of the average of the contributions from the blocks $H_i(\boldsymbol{\omega})$, which are assumed to be equal, and $F_i(\boldsymbol{\omega})$, which are not, since the image blocks have varying spectral content. Therefore the blur component should tend to dominate in the summation. It may be possible to remove this average image component, subtracting the average of a collection of representative unblurred images.

As an alternative to Equation (C.3), the *Power Cepstrum* may be used, which involves taking the FT of the log power spectra of the signals. Combined with the block-based method just described, the result is that a large spike will occur in the Cepstral domain wherever there was a periodic pattern of zeros in the original Fourier domain; the distance of this spike from the origin represents the spacing between the zeros and hence may be used to identify the parameters of the blurs. The large contribution from the blur term repeated in each block will dominate, and the high-frequency noise and spectrally averaged image content tend to be separated from the blur spike in the Cepstral domain. These methods have been extended to use the Power Bispectrum instead in [54] which shows improved performance in low SNR conditions. However, these methods are all limited to the parametric PSF models.

C.2 Zero Sheet Separation Algorithms

Lane and Bates [126] have shown that any signal $g(\mathbf{x})$, formed by multiple convolutions, in theory, is *automatically deconvolvable*, provided its dimension is greater than one. This

argument rests on the analytical properties of the Z-transform (ZT) of N-dimensional signals with finite support, which is necessarily zero on $(2N-2)$-dimensional hypersurfaces in a 2N-dimensional space.

The method assumes that there is no additive noise, i.e., the relation $G(z_1, z_2) = H(z_1, z_2)F(z_1, z_2)$ holds, where $F(z_1, z_2)$ and $H(z_1, z_2)$ are the 2-D ZTs of the original image and blur. Then the task is to separate the 2-D ZT of the blurred image, $G(z_1, z_2)$, into the two convolutive factors. As $G(z_1, z_2)$ is a polynomial in z_1, z_2, the solution is equivalent to factorizing the polynomial, i.e., identifying the zeros that belong to each component.

In order to do this, several assumptions are made: the convolutive factors (image and blur) should have compact (finite) support; the ZT of each factor should be zero on a single continuous surface, its *zero sheet*; that these zero sheets do not coalesce — they only intersect at discrete points (which it is suggested often holds in practice).

Zero sheets appear conceptually useful in analyzing the BID problem; however, in practice their applicability is limited. The main problem of this method is its high computational complexity and sensitivity to noise.

C.3 ARMA Parameter Estimation Algorithms

The estimation of the ARMA parameters in the model described in Section 4.4.2 can be done in many different ways; in addition to the Bayesian approaches already described, there also exist generalised cross-validation (GCV) and neural network-based algorithms, which are based on second-order statistics. Also, higher order statistics (HOS) approaches can be used for estimating the ARMA parameters for non-Gaussian models (it should be noted that with the second-order statistics algorithms the phase of the PSF can not be recovered, unless it is assumed that the PSF is minimum phase). The GCV algorithm will now be described.

Cross-validation is a parameter estimation method that partitions the observed data into two sets. Given some parameter values, the estimation set is used to form a prediction, and then the validation set is used to test the validity of this prediction. If the parameter values used were close to the correct ones, the validation criterion will be small. To make full use of the data, the criterion may be averaged across all the choices of estimation sets, using only one data element for each validation set. Hence the procedure is also referred to as the "leave one

out" algorithm. Prior to its use in BID, GCV was used in the estimation of the regularization parameter for classical restoration [70, 186].

When GCV is used to solve the BID problem [187], the original image $\hat{\mathbf{f}}$ is estimated using all but one of the pixels, $\mathbf{x}$, from the degraded image $\mathbf{g}$, for a particular value of the parameter set $\Theta = \{\mathbf{h}, \boldsymbol{\Omega}\}$ under test. The validation criterion is then to test the difference between the pixel $\mathbf{x}$ of $\mathbf{g}$ that was left out, and the corresponding reblurred pixel in $\mathbf{H}\hat{\mathbf{f}}$. This is averaged over all choices of $\mathbf{x}$. With the GCV criterion defined, a numerical search technique may be used to search for the parameter set that minimizes the criterion.

C.4 Nonparametric Deterministic Constraints Algorithms

This class of algorithms differs from the other joint blur and image identification methods in that they do not explicitly model the original image or the PSF with a stochastic or deterministic model. Instead, they typically use an iterative formulation to impose deterministic constraints on the unknowns at each step. These deterministic constraints may include nonnegativity, finite support, and energy bounds on the image, the blur, or both. These constraints are incorporated into an optimality criterion which is minimized with numerical iterative techniques.

Examples in the literature based on deterministic constraint principles include the iterative blind deconvolution (IBD) algorithm [7, 145, 216], McCallum's simulated annealing algorithm [138], the nonnegativity and support constraints with recursive image filtering (NAS-RIF) algorithm [117], and the blind superresolution algorithm [166], among others. Some of these methods will now be briefly described.

C.4.1 The Iterative Blind Deconvolution Algorithms

One of the early IBD algorithms is the one proposed by Ayers and Dainty [7]. In addition to nonnegativity and finite support, it uses Wiener-like constraints to estimate image and blur in the Fourier domain at each iteration. Beginning with an initial random PSF estimate $\widehat{h}_0(\mathbf{x})$ and image estimate $\widehat{f}_0(\mathbf{x})$, the following sequence defines the algorithm at iteration i:

1. Find $\widehat{F}_i(\mathbf{k})$, the DFT of $\widehat{f}_i(\mathbf{x})$.

2. Impose blur constraints in the Fourier domain (where $(.)^*$ denotes complex conjugation):

$$\widetilde{H}_i(\mathbf{k}) = \frac{G(\mathbf{k})\widehat{F}_i^*(\mathbf{k})}{\left|\widehat{F}_i(\mathbf{k})\right|^2 + \alpha / \left|\widehat{H}_{i-1}(\mathbf{k})\right|^2} \tag{C.4}$$

3. Find $\widetilde{h}_i(\mathbf{x})$, the IDFT of $\widetilde{H}_i(\mathbf{k})$.

4. Impose spatial domain positivity and finite support constraints on $\widetilde{h}_i(\mathbf{x})$ to give $\widehat{h}_i(\mathbf{x})$

5. Find $\widehat{H}_i(\mathbf{k})$, the DFT of $\widehat{h}_i(\mathbf{x})$.

6. Impose image constraints in the Fourier domain:

$$\widetilde{F}_i(\mathbf{k}) = \frac{G(\mathbf{k})\widehat{H}_i^*(\mathbf{k})}{\left|\widehat{H}_i(\mathbf{k})\right|^2 + \alpha / \left|\widehat{F}_i(\mathbf{k})\right|^2}, \tag{C.5}$$

7. Find $\widetilde{f}_i(\mathbf{x})$, the IDFT of $\widetilde{F}_i(\mathbf{k})$,

8. Impose spatial domain positivity and finite support constraints on $\widetilde{f}_i(\mathbf{x})$ to give $\widehat{f}_{i+1}(\mathbf{x})$.

9. Next iteration: set $i = i + 1$; go to step 1.

The real constant α represents the energy of the additive noise, and it has to be carefully chosen in order to obtain a reliable restoration. While this method is intuitively appealing, its convergence properties are undefined and tend to be highly sensitive to the initial guess [116]. Notice the similarity to the EM algorithm in the Fourier domain [107, 128]; however, the IBD algorithm is heuristically derived, and does not include estimation of the image and noise model parameters.

Further IBD-type algorithms using set-theoretic projection have been proposed by Yang *et al.* [232] and Lane [125], for the special case of astronomical speckle imaging.

C.4.2 The NAS-RIF Algorithms

The NAS-RIF method by Kundur and Hatzinakos [116, 117] is a similar method to IBD. It seeks to minimize a cost function at each step, by updating an FIR restoration filter, which is convolved with $\mathbf{g}$ to give an estimate $\hat{\mathbf{f}}$. The cost function is based on the same constraints as those in IBD, apart from that no assumptions are made on the PSF other than it having

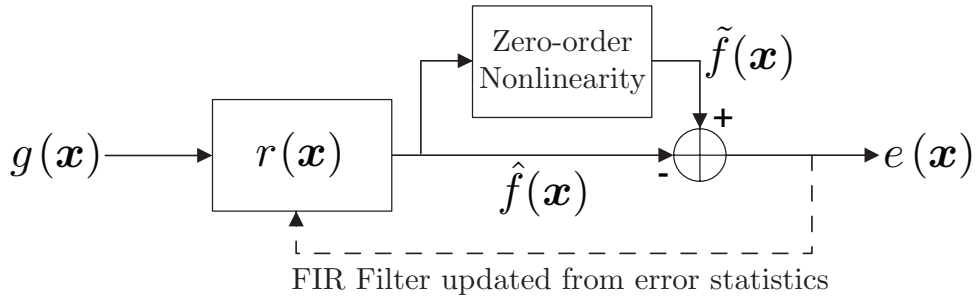


Figure C.1: *Blind image deconvolution based on higher order statistics.*

an inverse, both of which must be absolutely summable. The constraints are applied via the method of projection onto convex sets (POCS). NAS-RIF seems to have been fairly successful in its goals, with good convergence properties and reasonable quality restorations. However, it is only applicable to the class of images with finite support, i.e., it is entirely contained in the image frame on a uniform black background. This may include applications in medical imaging and astronomy, but prevents its widespread use with other natural images. Extensions to this method have appeared in [164, 168].

C.5 Nonparametric Algorithms based on Higher-Order Statistics

The principle of the HOS algorithms [40] – [225] is to use a nonlinear or non-Gaussian representation of the original image, allowing higher -order moments of the signal to be represented. These models have been typically applied when the image or its edges are modeled as sparse "spike-like" signals (for example, star fields).

In order to exploit HOS, an adaptive filter combined with a nonlinearity is used to restore the blurred image. The filter is updated to give a restored image that best fits the model. The adaptive filter structure for the HOS algorithms is shown in Figure C.1.

Note that $g(\mathbf{x})$ represents the degraded image, $\widehat{f}(\mathbf{x})$ represents the original image estimate, $\widetilde{f}(\mathbf{x})$ represents the output of the zero-order nonlinearity, $r(\mathbf{x})$ represents the restoration filter used to obtain the original image estimate, and $e(\mathbf{x})$ represents an error sample. The restoration filter $r(\mathbf{x})$ is optimized in order to minimize some cost function J that involves a sequence of error samples $e(\mathbf{x})$.

C.6 Total Least Squares (TLS)

Total Least Squares approaches are extensions of the standard least squares methods. The PSF is assumed to be the sum of a deterministic and a stochastic component, that is, $\mathbf{H} = \overline{\mathbf{H}} + \delta\mathbf{H}$. Using this, the degradation model in Equation (2.6) can be expressed as

$$\overline{\mathbf{g}} + \delta\mathbf{g} = (\overline{\mathbf{H}} + \delta\mathbf{H})\mathbf{f} + \boldsymbol{w}, \tag{C.6}$$

where $\overline{\mathbf{g}}$ and $\delta\mathbf{g}$ are the deterministic and stochastic components of the observation $\mathbf{g}$, respectively. The problem is formulated as a minimization of $\delta\mathbf{g}$ and $\delta\mathbf{H}$ subject to Equation (C.6). Generally, regularized constrained total least squares (RCTLS) filters [139] are applied to find the minimum values, where the matrices are assumed to have a special form, such as BTTB. In addition to the TLS solutions already presented that use the hierarchical Bayesian framework [71], linear algebra-based solutions are applied in [137].

C.7 Learning-Based Algorithms

Learning-based algorithms for image restoration and blind image restoration have been recently proposed [159, 173]. The basic idea with such an approach is that the prior knowledge required for solving various (inverse) problems can be learned from training data, i.e., a set of prototype images belonging to the same (statistical) class of images to the ones processed.

Original images and their degraded versions by the known degradation operator (restoration problem) are used for designing vector quantizer (VQ) codebooks. The codevectors are designed using the blurred images. For each such vector, the high frequency information obtained from the original images is also available. During restoration, the high-frequency information of a given degraded image is estimated from its low-frequency information based on the codebooks. For the BID problem, a number of codebooks are designed corresponding to various versions of the blurring function. Given a noisy and blurred image, one of the codebooks is chosen based on a similarity measure, therefore providing the identification of the blur. To make the restoration process computationally efficient, principal component analysis (PCA) and VQ-Nearest Neighborhood approaches are utilized in [159].

C.8 Methods for Spatially Varying Degradation

Blind deconvolution in case of spatially varying (SV) degradation is a more difficult problem than the spatially invariant case. The blur is generally assumed to be varying smoothly or piecewise-smoothly, and the variation to be slow in the spatial domain. The standard EM procedure has been extended to use sectioned methods where the image is divided into blocks [84, 119]. A hierarchical sliding window approach with the local Fourier transform is employed in [169]. In [182], SV PSF identification for a known image is considered using an MRF model for the parameterization of the SV blur. In [157], the authors propose a preconditioning method for speeding up SV restoration applied via a nonnstationary convolution matrix consisting of linearly interpolated kernels. Some of the spatially invariant methods described in previous sections are also extended to the SV blur case, for example, a parameterized piecewise-smooth degradation model is used to extend the anisotropic regularization-based restoration method in [236].

C.9 Multichannel Methods

Multichannel images are typically acquired using an imaging system with multiple sensors, multiple time instants, or multiple frequency bands. Examples of multichannel images include multispectral images, where different channels present different frequency bands, wave radiometric images, and image sequences, such as video. Reviews of classical and blind multichannel restoration methods are presented in [69, 207]. Multichannel methods can be classified into two approaches, single input multiple output (SIMO) and multiple input multiple output (MIMO).

The EM approach in [128] is combined with the MIMO restoration method in [111] to obtain a blind multichannel method in [213], where cross-channel blurs are also taken into account. The SIMO multichannel restoration problem is addressed in [181] by an ARMA model, where the likelihood function maximisation is performed using a steepest descent algorithm.

Another class of algorithms have been developed using greatest common divisor (GCD) methods. Under the *relative coprimeness* condition — the channel PSFs share no common factor other than a scalar — the GCD of the outputs of the channels will be the original image. By exploiting the commutativity of the convolution operator, a matrix equation is formed in-

volving channel outputs and PSFs. Then, as an extension of 1-D blind equalization methods, eigenstructure properties of the matrices may be used to directly estimate the PSFs and the original image. Perfect restoration algorithms for noise-free degradation with FIR PSFs have been proposed in [90] and [89]. A similar approach with a direct estimation method has been proposed in [77]. The noise is a major problem for these approaches. To deal with noise amplification, direct vector-space methods have been proposed in [171] and [170], and sufficiency conditions are derived for exact restoration in [172]. Rav-Acha and Peleg [184] have proposed a multichannel method with an *a priori* blur identification method via an exhaustive search, and a coprimeness condition is imposed on the channel model.

Recently an extension of [90] and [89] has been proposed in [205], that exploits anisotropic regularization priors mentioned in Section B.4. As well as the standard TV form, a more advanced Mumford-Shah regularization term is also used. Very good restoration results are achieved even in low SNR conditions. This method is extended to deal with unknown PSF support and global translational motion in [206], where a MAP formulation is utilized.

Observe that there is an inherent advantage with multichannel methods in the amount of information available to aid both blur identification and image restoration. The coprimeness condition ensures that the problem becomes less ill-posed; therefore we should expect better results than in single-channel methods.

Appendix D

Marginalisation of image model parameters

The integrals that were needed to marginalise the nuisance AR and variance parameters from the MMAP model posterior expression $p(\mathbf{h}, \boldsymbol{\alpha}_f, \boldsymbol{\sigma}_v \,|\, \mathbf{g})$, (7.16), repeated here:

$$\frac{p(\boldsymbol{\sigma}_v)\det|\mathbf{R}_v|^{-\frac{1}{2}}}{(2\pi)^{\frac{(L_+ + L_{\alpha_f})}{2}}\det|\mathbf{H}|\det|\mathbf{Q}_v|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}\left(\boldsymbol{\alpha}_f^T\{\hat{\mathbf{F}}^T\mathbf{Q}_v^{-1}\hat{\mathbf{F}}+\delta_{\alpha_f}^{-1}\mathbf{R}_v^{-1}\}\boldsymbol{\alpha}_f - 2\boldsymbol{\alpha}_f^T\hat{\mathbf{F}}^T\mathbf{Q}_v^{-1}\hat{\mathbf{f}}+\hat{\mathbf{f}}^T\mathbf{Q}_v^{-1}\hat{\mathbf{f}}\right)\right]$$

are detailed in the following sections.

D.1 Marginalisation of AR parameters

Using the standard Gaussian integral (7.5), with $\Gamma = \hat{\mathbf{F}}^T\mathbf{Q}_v^{-1}\hat{\mathbf{F}}+\delta_{\alpha_f}^{-1}\mathbf{R}_v^{-1}$, $\boldsymbol{\beta} = \hat{\mathbf{F}}^T\mathbf{Q}_v^{-1}\hat{\mathbf{f}}$ and $\alpha = \hat{\mathbf{f}}^T\mathbf{Q}_v^{-1}\hat{\mathbf{f}}$, the posterior may be marginalised to remove the AR parameter vector, $\boldsymbol{\alpha}_f$, as follows:

$$\begin{aligned} p(\mathbf{h},\boldsymbol{\sigma}_v|\mathbf{g}) &= \int_{\mathbb{R}^{L_{\alpha_f}\times 1}}\cdots\int p(\mathbf{h},\boldsymbol{\alpha}_f,\boldsymbol{\sigma}_v\,|\,\mathbf{g})\,d\boldsymbol{\alpha}_f \\ &\propto \frac{p(\boldsymbol{\sigma}_v)\det|\mathbf{R}_v|^{-\frac{1}{2}}}{(2\pi)^{\frac{(L_+ + L_{\alpha_f})}{2}}\det|\mathbf{H}|\det|\mathbf{Q}_v|^{\frac{1}{2}}}\cdot\frac{(2\pi)^{L_{\alpha_f}/2}}{\det\left|(\hat{\mathbf{F}}^T\mathbf{Q}_v^{-1}\hat{\mathbf{F}}+\delta_{\alpha_f}^{-1}\mathbf{R}_v^{-1})\right|^{\frac{1}{2}}} \\ &\quad\cdot\exp\left[-\frac{1}{2}\left(\hat{\mathbf{f}}^T\mathbf{Q}_v^{-1}\hat{\mathbf{f}}-(\hat{\mathbf{F}}^T\mathbf{Q}_v^{-1}\hat{\mathbf{f}})^T(\hat{\mathbf{F}}^T\mathbf{Q}_v^{-1}\hat{\mathbf{F}}+\delta_{\alpha_f}^{-1}\mathbf{R}_v^{-1})^{-1}(\hat{\mathbf{F}}^T\mathbf{Q}_v^{-1}\hat{\mathbf{f}})\right)\right], \end{aligned} \tag{D.1}$$

noting the terms not dependant on $\boldsymbol{\alpha}_f$ may be regarded as constants and taken outside the integral. Now the two models for excitation variance may be considered (§D.2 and §D.3).

D.2 Marginalisation of single variance parameter for whole image

Simplifying using the assumption that $\mathbf{Q}_\nu = \sigma_\nu^2 \mathrm{I}_{\mathrm{L}_+} \Rightarrow \det|\mathrm{Q}_\nu|^{\frac{1}{2}} = \sigma_\nu^{\mathrm{L}_+}$, thus $\mathbf{R}_\nu = \sigma_\nu^2 \mathrm{I}_{\mathrm{L}_{a_f}} \Rightarrow \det|\mathbf{R}_\nu|^{\frac{1}{2}} = \sigma_\nu^{\mathrm{L}_{a_f}}$, and substituting the prior from (7.13):

$$\begin{aligned} p(\mathbf{h}, \sigma_\nu^2|\mathbf{g}) \propto & \frac{\sigma_\nu^{-\mathrm{L}_{a_f}} \det\left|\sigma_\nu^{-2}(\hat{\mathbf{F}}^\mathsf{T}\hat{\mathbf{F}} + \delta_{a_f}^{-1}\mathbf{I}_{\mathrm{L}_{a_f}})\right|^{-\frac{1}{2}} \beta^\alpha}{(2\pi)^{\frac{(\mathrm{L}_+)}{2}} \det|\mathbf{H}| \sigma_\nu^{\mathrm{L}_+} \sigma_\nu^{(2\alpha+2)} \Gamma(\alpha)} \\ & \cdot \exp\left[-\frac{1}{2\sigma_\nu^2}\left(\hat{\mathbf{f}}^\mathsf{T}\hat{\mathbf{f}} - \hat{\mathbf{f}}^\mathsf{T}\hat{\mathbf{F}}(\hat{\mathbf{F}}^\mathsf{T}\hat{\mathbf{F}} + \delta_{a_f}^{-1}\mathbf{I}_{\mathrm{L}_{a_f}})^{-1}\hat{\mathbf{F}}^\mathsf{T}\hat{\mathbf{f}} + 2\beta\right)\right] \end{aligned}$$

and using the standard gamma integral:

$$\frac{\Gamma(\epsilon)}{Q^\epsilon} = \int_0^\infty \gamma^{\epsilon-1} e^{-Q\gamma}\, d\gamma\,, \tag{D.2}$$

the excitation variance parameter σ_ν^2 may too be marginalised. This is achieved by change of variables; letting $\gamma = \sigma_\nu^{-2} \Rightarrow \frac{d\sigma_\nu}{d\gamma} = -\frac{1}{2}\gamma^{-\frac{3}{2}}$, we have:

$$p(\mathbf{h}|\mathbf{g}) = \int_0^\infty p(\mathbf{h}, \sigma_\nu^2|\mathbf{g})\, d\sigma_\nu = \int_\infty^0 p(\mathbf{h}, \sigma_\nu^2|\mathbf{g}) \cdot -\frac{1}{2}\gamma^{-\frac{3}{2}}\, d\gamma \tag{D.3}$$

$$\propto \frac{\det\left|\hat{\mathbf{F}}^\mathsf{T}\hat{\mathbf{F}} + \delta_{a_f}^{-1}\mathbf{I}_{\mathrm{L}_{a_f}}\right|^{-\frac{1}{2}}}{(2\pi)^{\frac{(\mathrm{L}_+)}{2}} \det|\mathbf{H}|} \cdot \int_0^\infty \gamma^{\frac{\mathrm{L}_+ + 2\alpha + 2}{2}} \cdot \exp\left[-Q\gamma\right] \cdot \frac{1}{2}\gamma^{-\frac{3}{2}}\, d\gamma, \tag{D.4}$$

where $Q = \frac{1}{2}\left(\hat{\mathbf{f}}^\mathsf{T}\hat{\mathbf{f}} - \hat{\mathbf{f}}^\mathsf{T}\hat{\mathbf{F}}(\hat{\mathbf{F}}^\mathsf{T}\hat{\mathbf{F}} + \delta_{a_f}^{-1}\mathbf{I}_{\mathrm{L}_{a_f}})^{-1}\hat{\mathbf{F}}^\mathsf{T}\hat{\mathbf{f}} + 2\beta\right)$. Now comparing (D.4) with the R.H.S. of (D.2), it is seen that $\epsilon - 1 = \frac{\mathrm{L}_+ - 1 + 2\alpha}{2} \Rightarrow \epsilon = \frac{\mathrm{L}_+ + 1 + 2\alpha}{2}$,

$$\therefore p\,(\mathbf{h}\,|\,\mathbf{g}\,) \propto \frac{1}{2}\frac{\Gamma(\mathrm{L}_+)\det\left|\hat{\mathbf{F}}^\mathsf{T}\hat{\mathbf{F}} + \delta_{a_f}^{-1}\mathbf{I}_{\mathrm{L}_{a_f}}\right|^{-\frac{1}{2}}}{(2\pi)^{\frac{(\mathrm{L}_+)}{2}} \det|\mathbf{H}|} \left(\hat{\mathbf{f}}^\mathsf{T}\hat{\mathbf{f}} - \hat{\mathbf{f}}^\mathsf{T}\hat{\mathbf{F}}(\hat{\mathbf{F}}^\mathsf{T}\hat{\mathbf{F}} + \delta_{a_f}^{-1}\mathbf{I}_{\mathrm{L}_{a_f}})^{-1}\hat{\mathbf{F}}^\mathsf{T}\hat{\mathbf{f}} + 2\beta\right)^{-\frac{\mathrm{L}_+ + 1 + 2\alpha}{2}}. \tag{D.5}$$

D.3 Marginalisation of individual variance parameter for each block

In the more general case when each block has its own variance, we have $\mathbf{Q}_{\nu'} = \mathrm{diag}([\sigma_{\nu_1}^2 \mathbf{1}_{\mathrm{L}_1}^\mathsf{T} \cdots \sigma_{\nu_R}^2 \mathbf{1}_{\mathrm{L}_R}^\mathsf{T}]^\mathsf{T})$, and $\mathbf{Q}_\nu = \mathbf{P}\mathbf{Q}_{\nu'}\mathbf{P}^\mathsf{T}$.

As an aside, note that in the case we have $M \times N$ regular rectanguar blocks for the regions, each of size $m \times n$, we may use the Kronecker product, $\otimes$, and the rowwise lexicographic ordering operator applied to a matrix, vec, to write

$$\mathbf{Q}_\nu = \mathrm{diag}\left(\mathrm{vec}(\mathbf{1}_{\frac{m}{M}}^{\mathsf{T}} \otimes \boldsymbol{\sigma} \otimes \mathbf{1}_{\frac{n}{N}}^{\mathsf{T}})\right) \tag{D.6}$$

where $\boldsymbol{\sigma} = [\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2 \cdots \boldsymbol{\sigma}_M]^{\mathsf{T}}$, $\boldsymbol{\sigma}_I = [\sigma_{I,1}, \sigma_{I,2} \cdots \sigma_{I,N}]^{\mathsf{T}}$, and the equivalent notation $\{\sigma_{\nu_r}\}_{r\in\{1\cdots R\}} \equiv \{\sigma_{I,J}\}_{\{I,J\}\in\{1\cdots M,1\cdots N\}}$ is used. We may also describe the diagonal covariance matrix for the AR parameters, $\mathbf{R}_\nu$, of size $L_{a_f} \times L_{a_f}$, in a similar way:

$$\mathbf{R}_\nu = \mathrm{diag}\left(\mathrm{vec}(\boldsymbol{\sigma} \otimes \mathbf{1}_{L_{a_{f_r}}}^{\mathsf{T}})\right). \tag{D.7}$$

However, because in any case the distribution of $\boldsymbol{\sigma}_\nu$ and the marginal distribution in (D.1) are separable over the regions, we can simply expand the PDF into a product over each region, and marginalise each separately. The determinant terms involving $\hat{\mathbf{F}}$ may be expanded by noting that the determinant of a block-diagonal matrix is a product of the blocks, and its inverse is a block diagonal matrix containing the inverse of each block. Thus incorporating the prior $p(\boldsymbol{\sigma}_\nu \mid \boldsymbol{\alpha}_\nu, \boldsymbol{\beta}_\nu) = \prod_{r\in\{1\cdots R\}} \mathcal{IG}(\sigma_{\nu_r}^2 \mid \alpha_{\nu_r}, \beta_{\nu_r})$ with (D.1) we obtain:

$$\begin{aligned} p(\mathbf{h}, \boldsymbol{\sigma}_\nu \mid \mathbf{g}) \propto{}& \frac{\left(\prod_{r\in\{1\cdots R\}} p(\sigma_{\nu_r}^2 \mid \alpha_{\nu_r}, \beta_{\nu_r})\right) \det\left|(\hat{\mathbf{F}}^{\mathsf{T}}\mathbf{Q}_\nu^{-1}\hat{\mathbf{F}} + \delta_{a_f}^{-1}\mathbf{R}_\nu^{-1})\right|^{-\frac{1}{2}}}{\det|\mathbf{H}|\det|\mathbf{Q}_\nu|^{\frac{1}{2}}\det|\mathbf{R}_\nu|^{\frac{1}{2}}} \\ &\cdot \exp\left[-\frac{1}{2}\left(\hat{\mathbf{f}}^{\mathsf{T}}\mathbf{Q}_\nu^{-1}\hat{\mathbf{f}} - (\hat{\mathbf{F}}^{\mathsf{T}}\mathbf{Q}_\nu^{-1}\hat{\mathbf{f}})^{\mathsf{T}}(\hat{\mathbf{F}}^{\mathsf{T}}\mathbf{Q}_\nu^{-1}\hat{\mathbf{F}} + \delta_{a_f}^{-1}\mathbf{R}_\nu^{-1})^{-1}(\hat{\mathbf{F}}^{\mathsf{T}}\mathbf{Q}_\nu^{-1}\hat{\mathbf{f}})\right)\right], \end{aligned} \tag{D.8}$$

$$\begin{aligned} \propto{}& \frac{\left(\prod_{r\in\{1\cdots R\}} (\sigma_{\nu_r}^2)^{-(\alpha_{\nu_r}+1)} \exp\left[-\frac{\beta_{\nu_r}}{\sigma_{\nu_r}^2}\right]\right)\left(\prod_{r\in\{1\cdots R\}} \sigma_{\nu_r}^{L_{a_{f_r}}} \det\left|(\hat{\mathbf{F}}_r^{\mathsf{T}}\hat{\mathbf{F}}_r + \delta_{a_f}^{-1}\mathbf{I}_{L_{a_{f_r}}})\right|^{-\frac{1}{2}}\right)}{\det|\mathbf{H}| \prod_{r\in\{1\cdots R\}} \sigma_{\nu_r}^{L_r}\sigma_{\nu_r}^{L_{a_{f_r}}}} \\ &\cdot \exp\left[-\frac{1}{2}\sum_{r\in\{1\cdots R\}} \frac{1}{\sigma_{\nu_r}^2}\left(\hat{\mathbf{f}_r}^{\mathsf{T}}\hat{\mathbf{f}_r} - \hat{\mathbf{f}_r}^{\mathsf{T}}\hat{\mathbf{F}}_r(\hat{\mathbf{F}}_r^{\mathsf{T}}\hat{\mathbf{F}}_r + \delta_{a_f}^{-1}\mathbf{I}_{L_{a_{f_r}}})^{-1}\hat{\mathbf{F}}_r^{\mathsf{T}}\hat{\mathbf{f}_r}\right)\right], \end{aligned} \tag{D.9}$$

Then integrating out all the $\sigma_{\nu_r}^2$ (note that terms not containing $\sigma_{\nu_r}^2$ are taken outside the r^{th} integral; thus the terms are completely separable, and the Jacobians are just the transformation

of each variable individually — as before $\gamma_r = \sigma_{\nu_r}^{-2} \Rightarrow \frac{d\sigma_{\nu_r}}{d\gamma_r} = -\frac{1}{2}\gamma_r^{-\frac{3}{2}}$):

$$p(\mathbf{h}|\mathbf{g}) = \int_0^\infty \cdots \int p(\mathbf{h}, \{\sigma_{\nu_r}^2\}|\hat{\mathbf{f}_r})\, d\sigma_{\nu_1} \cdots d\sigma_{\nu_R} \tag{D.10}$$

$$= \int_\infty^0 \cdots \int p(\mathbf{h}, \{\sigma_{\nu_r}^2\}|\hat{\mathbf{f}_r}) \cdot \left\{ \prod_{r\in\{1\cdots R\}} \left(-\frac{1}{2}\gamma_r^{-\frac{3}{2}}\right) \right\} d\gamma_1 \cdots d\gamma_R \tag{D.11}$$

$$\propto \frac{1}{\det|\mathbf{H}|} \prod_{r\in\{1\cdots R\}} \left(\det\left|\hat{\mathbf{F}}_r^T\hat{\mathbf{F}}_r + \delta_{a_f}^{-1}\mathbf{I}_{L_{a_{f_r}}}\right|^{-\frac{1}{2}} \cdot \int_0^\infty \gamma_r^{\frac{L_r+2\alpha_{\nu_r}+2}{2}} \cdot \exp\left[-Q_r\gamma_r\right] \cdot \frac{1}{2}\gamma_r^{-\frac{3}{2}}\, d\gamma_r \right) \tag{D.12}$$

where $Q_r = \frac{1}{2}\left(\hat{\mathbf{f}_r}^T\hat{\mathbf{f}_r} - \hat{\mathbf{f}_r}^T\hat{\mathbf{F}}_r(\hat{\mathbf{F}}_r^T\hat{\mathbf{F}}_r + \delta_{a_f}^{-1}\mathbf{I}_{L_{a_{f_r}}})^{-1}\hat{\mathbf{F}}_r^T\hat{\mathbf{f}_r} + 2\beta_{\nu_r}\right)$

Now with reference to the gamma integral (D.2), we have $\epsilon_r - 1 = \frac{L_r-1+2\alpha_{\nu_r}}{2} \Rightarrow \epsilon_r = \frac{L_r+1+2\alpha_{\nu_r}}{2}$, and we obtain the result:

$$p(\mathbf{h}|\mathbf{g}) \propto \frac{1}{\det|\mathbf{H}|} \prod_{r\in\{1\cdots R\}} \left(\frac{\left(\hat{\mathbf{f}_r}^T\hat{\mathbf{f}_r} - \hat{\mathbf{f}_r}^T\hat{\mathbf{F}}_r(\hat{\mathbf{F}}_r^T\hat{\mathbf{F}}_r + \delta_{a_f}^{-1}\mathbf{I}_{L_{a_{f_r}}})^{-1}\hat{\mathbf{F}}_r^T\hat{\mathbf{f}_r} + 2\beta_{\nu_r}\right)^{-\frac{L_r+1+2\alpha_{\nu_r}}{2}}}{\det\left|\hat{\mathbf{F}}_r^T\hat{\mathbf{F}}_r + \delta_{a_f}^{-1}\mathbf{I}_{L_{a_{f_r}}}\right|^{\frac{1}{2}}} \right) \tag{D.13}$$

Appendix E
Marginalisation of image boundary

E.1 Proof via Marginalisation

In this section we use the Gaussian integral to marginalise the boundary. Firstly expanding the likelihood,

$$\mathrm{p}\left(\mathbf{g} \,|\, \mathbf{f}, \mathbf{H}, \mathbf{Q}_w\right) = \mathcal{N}\left(\mathbf{g} \,|\, \mathbf{H}\mathbf{f}, \mathbf{Q}_w\right) \tag{E.1}$$

$$= \mathcal{N}\left(\mathbf{g} \,|\, \mathbf{H}_1\mathbf{f}_1 + \mathbf{H}_0\mathbf{f}_0, \mathbf{Q}_w\right) \tag{E.2}$$

$$\propto \exp\Big(-\frac{1}{2}\left[\mathbf{f}^{\mathsf{T}}(\mathbf{H}^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{H})\mathbf{f} - 2\mathbf{f}^{\mathsf{T}}(\mathbf{H}^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{g}) + \mathbf{g}^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{g}\right]\Big) \tag{E.3}$$

$$\begin{aligned} = \exp\Bigg[-\frac{1}{2}\Bigg(&\left[\begin{array}{c|c} \mathbf{f}_1 & \mathbf{f}_0 \end{array}\right] \left[\begin{array}{c|c} \mathbf{H}_1\mathbf{Q}_w^{-1}\mathbf{H}_1 & \mathbf{H}_1\mathbf{Q}_w^{-1}\mathbf{H}_0 \\ \hline \mathbf{H}_0\mathbf{Q}_w^{-1}\mathbf{H}_1 & \mathbf{H}_0\mathbf{Q}_w^{-1}\mathbf{H}_0 \end{array}\right] \left[\begin{array}{c} \mathbf{f}_1 \\ \hline \mathbf{f}_0 \end{array}\right] \\ &- 2\left[\begin{array}{c|c} \mathbf{f}_1 & \mathbf{f}_0 \end{array}\right] \left[\begin{array}{c} \mathbf{H}_1 \\ \hline \mathbf{H}_0 \end{array}\right] \mathbf{Q}_w^{-1}\mathbf{g} + \mathbf{g}^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{g}\Bigg)\Bigg] \end{aligned} \tag{E.4}$$

$$\begin{aligned} = \exp\Bigg[-\frac{1}{2}\Big(&\mathbf{f}_1^{\mathsf{T}}(\mathbf{H}_1^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{H}_1)\mathbf{f}_1 + 2\mathbf{f}_0^{\mathsf{T}}(\mathbf{H}_0^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{H}_1)\mathbf{f}_1 + \mathbf{f}_0^{\mathsf{T}}(\mathbf{H}_0^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{H}_0)\mathbf{f}_0 \\ &- 2\mathbf{f}_1^{\mathsf{T}}(\mathbf{H}_1^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{g}) - 2\mathbf{f}_0^{\mathsf{T}}(\mathbf{H}_0^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{g}) + \mathbf{g}^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{g}\Big)\Bigg] \end{aligned} \tag{E.5}$$

$$\begin{aligned} = \exp\Bigg[-\frac{1}{2}\Big(&\mathbf{f}_1^{\mathsf{T}}(\mathbf{H}_1^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{H}_1)\mathbf{f}_1 - 2\mathbf{f}_1^{\mathsf{T}}\left(\mathbf{H}_1^{\mathsf{T}}\mathbf{Q}_w^{-1}(\mathbf{g} - \mathbf{H}_0\mathbf{f}_0)\right) \\ &+ \left(\mathbf{g}^{\mathsf{T}} - \mathbf{H}_0\mathbf{f}_0\right)^{\mathsf{T}}\mathbf{Q}_w^{-1}\left(\mathbf{g}^{\mathsf{T}} - \mathbf{H}_0\mathbf{f}_0\right)\Big)\Bigg] \end{aligned} \tag{E.6}$$

and the prior for $\mathbf{f}$,

$$\mathrm{p}\left(\mathbf{f} \,\middle|\, \bar{\mathbf{f}}_+, \boldsymbol{\Sigma}_{\mathrm{f}_+}\right) = \mathcal{N}\left(\mathbf{f} \,\middle|\, \bar{\mathbf{f}}_+, \boldsymbol{\Sigma}_{\mathrm{f}_+}\right) \tag{E.7}$$

$$= \mathcal{N}\left(\left[\begin{array}{c} \mathbf{f}_1 \\ \hline \mathbf{f}_0 \end{array}\right] \,\middle|\, \left[\begin{array}{c} 0 \\ \hline \bar{\mathbf{f}}_0 \end{array}\right], \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{\mathrm{f}_1}^{-1} & \boldsymbol{\Sigma}_{\mathrm{f}_1,\mathrm{f}_0}^{-1} \\ \hline \boldsymbol{\Sigma}_{\mathrm{f}_1,\mathrm{f}_0}^{-\mathsf{T}} & \boldsymbol{\Sigma}_{\mathrm{f}_0}^{-1} \end{array}\right]^{-1}\right), \tag{E.8}$$

$$\propto \exp\Big(-\frac{1}{2}\left[(\mathbf{f} - \bar{\mathbf{f}}_+)^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathrm{f}_+}^{-1}(\mathbf{f} - \bar{\mathbf{f}}_+)\right]\Big) \tag{E.9}$$

$$= \exp\Big[-\frac{1}{2}\Big(\Big(\mathbf{f}_1^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{f}_1}^{-1}\mathbf{f}_1 + 2\mathbf{f}_1^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{f}_1,\mathbf{f}_0}^{-1}\mathbf{f}_0 + \mathbf{f}_0^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{f}_0}^{-1}\mathbf{f}_0\Big) - 2\Big(\bar{\mathbf{f}}_0^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{f}_0}^{-1}\mathbf{f}_0 + \mathbf{f}_1^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{f}_1,\mathbf{f}_0}^{-1}\bar{\mathbf{f}}_0\Big) + \bar{\mathbf{f}}_0^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{f}_0}^{-1}\bar{\mathbf{f}}_0^{\mathsf{T}}\Big)\Big], \tag{E.10}$$

the integral of (8.9) is found as

$$\mathrm{p}\left(\mathbf{f}_1 \mid \mathbf{g}, \cdot\right) \propto \int \mathrm{p}\left(\mathbf{g} \mid \mathbf{f}, \mathbf{H}, \mathbf{Q}_w\right) \mathrm{p}\left(\mathbf{f} \,\middle|\, \bar{\mathbf{f}}_+, \boldsymbol{\Sigma}_{\mathbf{f}_+}\right) \mathrm{d}\mathbf{f}_0 \tag{E.11}$$

$$\propto \int \exp\Big[-\frac{1}{2}\Big(\mathbf{f}_0^{\mathsf{T}}\overbrace{(\mathbf{H}_0^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{H}_0 + \boldsymbol{\Sigma}_{\mathbf{f}_0}^{-1})}^{\Gamma}\mathbf{f}_0 - 2\overbrace{\Big((\mathbf{g}^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{H}_0 + \bar{\mathbf{f}}_0^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{f}_0}^{-1}) - \mathbf{f}_1^{\mathsf{T}}(\mathbf{H}_1^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{H}_0 + \boldsymbol{\Sigma}_{\mathbf{f}_1,\mathbf{f}_0}^{-1})\Big)}^{\beta^{\mathsf{T}}}\mathbf{f}_0 + \overbrace{\mathbf{f}_1^{\mathsf{T}}(\mathbf{H}_1^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{H}_1 + \boldsymbol{\Sigma}_{\mathbf{f}_1}^{-1})\mathbf{f}_1 - 2\mathbf{f}_1^{\mathsf{T}}(\mathbf{H}_1^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{g} + \boldsymbol{\Sigma}_{\mathbf{f}_1,\mathbf{f}_0}^{-1}\bar{\mathbf{f}}_0) + \mathbf{g}^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{g} + \bar{\mathbf{f}}_0^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{f}_0}^{-1}\bar{\mathbf{f}}_0}^{\alpha}\Big)\Big]\mathrm{d}\mathbf{f}_0 \tag{E.12}$$

which, using the Gaussian integral (7.5) and defining $\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_0} = \Gamma^{-1} = (\mathbf{H}_0^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{H}_0 + \boldsymbol{\Sigma}_{\mathbf{f}_0}^{-1})^{-1}$ and $\boldsymbol{\Sigma}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}}^{-1} = (\mathbf{H}_1^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{H}_0 + \boldsymbol{\Sigma}_{\mathbf{f}_1,\mathbf{f}_0}^{-1})$, results in:

$$\begin{aligned}\mathrm{p}\left(\mathbf{f}_1 \mid \mathbf{g}, \cdot\right) \propto \exp\Big[-\frac{1}{2}\Big(&\mathbf{f}_1^{\mathsf{T}}\Big(\mathbf{H}_1^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{H}_1 + \boldsymbol{\Sigma}_{\mathbf{f}_1}^{-1} - \boldsymbol{\Sigma}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}}^{-1}\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_0}\boldsymbol{\Sigma}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}}^{-\mathsf{T}}\Big)\mathbf{f}_1 \\ &+ 2\mathbf{f}_1^{\mathsf{T}}\boldsymbol{\Sigma}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}}^{-1}\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_0}(\mathbf{H}_0^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{g} + \boldsymbol{\Sigma}_{\mathbf{f}_0}^{-\mathsf{T}}\bar{\mathbf{f}}_0) \\ &- 2\mathbf{f}_1^{\mathsf{T}}\Big(\mathbf{H}_1^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{g} + \boldsymbol{\Sigma}_{\mathbf{f}_1,\mathbf{f}_0}^{-1}\bar{\mathbf{f}}_0^{\mathsf{T}}\Big) + \{\text{const. terms independent of } \mathbf{f}_1\}\Big)\Big]\end{aligned} \tag{E.13}$$

which is a Gaussian

$$\mathrm{p}\left(\mathbf{f}_1 \mid \mathbf{g}\right) \propto \mathcal{N}\left(\mathbf{f}_1 \,\middle|\, \boldsymbol{\mu}_{\hat{\mathbf{f}}}, \boldsymbol{\Sigma}_{\widehat{\mathbf{f}}}\right) \tag{E.14}$$

$$\propto \exp\Big(-\frac{1}{2}\big[\hat{\mathbf{f}}^{\mathsf{T}}\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}}^{-1}\hat{\mathbf{f}} - 2\hat{\mathbf{f}}^{\mathsf{T}}\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}}^{-1}\boldsymbol{\mu}_{\hat{\mathbf{f}}} + \boldsymbol{\mu}_{\hat{\mathbf{f}}}^{\mathsf{T}}\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}}^{-1}\boldsymbol{\mu}_{\hat{\mathbf{f}}}\big]\Big). \tag{E.15}$$

with parameters

$$\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}}^{-1} = \mathbf{H}_1^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{H}_1 + \boldsymbol{\Sigma}_{\mathbf{f}_1}^{-1} - \boldsymbol{\Sigma}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}}^{-1}\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_0}\boldsymbol{\Sigma}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}}^{-\mathsf{T}} \tag{E.16}$$

$$\boldsymbol{\mu}_{\hat{\mathbf{f}}} = \boldsymbol{\Sigma}_{\widehat{\mathbf{f}}}\Big(\Big(\mathbf{H}_1^{\mathsf{T}}\mathbf{Q}_w^{-1} - \boldsymbol{\Sigma}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}}^{-1}\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_0}\mathbf{H}_0^{\mathsf{T}}\mathbf{Q}_w^{-1}\Big)\mathbf{g} + \Big(\boldsymbol{\Sigma}_{\mathbf{f}_1,\mathbf{f}_0}^{-1} - \boldsymbol{\Sigma}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}}^{-1}\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_0}\boldsymbol{\Sigma}_{\mathbf{f}_0}^{-\mathsf{T}}\Big)\bar{\mathbf{f}}_0\Big). \tag{E.17}$$

E.2 Proof via Block Matrix inversion

An alternative slightly quicker derivation is found making use of the block-matrix inversion formula.

Firstly recall the posterior in (8.6) for $\mathbf{f}$ is $p(\mathbf{f} \mid \mathbf{g}) \sim \mathcal{N}\left(\mathbf{f} \,\middle|\, \boldsymbol{\mu}_{\hat{\mathbf{f}}_+}, \boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_+}\right)$. We can expand $\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_+}^{-1}$ as:

$$\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_+}^{-1} = \mathbf{H}^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{H} + \boldsymbol{\Sigma}_{\mathbf{f}_+}^{-1} \tag{E.18}$$

$$= \left[\begin{array}{c|c} \mathbf{H}_1^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{H}_1 + \boldsymbol{\Sigma}_{\mathbf{f}_1}^{-1} & \mathbf{H}_1^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{H}_0 + \boldsymbol{\Sigma}_{\mathbf{f}_1,\mathbf{f}_0}^{-1} \\ \hline \mathbf{H}_0^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{H}_1 + \boldsymbol{\Sigma}_{\mathbf{f}_1,\mathbf{f}_0}^{-\mathsf{T}} & \mathbf{H}_0^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{H}_0 + \boldsymbol{\Sigma}_{\mathbf{f}_0}^{-1} \end{array}\right] = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_1}^{-1} & \boldsymbol{\Sigma}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}}^{-1} \\ \hline \boldsymbol{\Sigma}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}}^{-\mathsf{T}} & \boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_0}^{-1} \end{array}\right] \tag{E.19}$$

This is inverted using the following identity for a symmetric positive definite matrix [101]:

$$\left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{B}^{\mathsf{T}} & \mathbf{C} \end{array}\right]^{-1} = \left[\begin{array}{c|c} \widetilde{\mathbf{C}}^{-1} & -\widetilde{\mathbf{C}}^{-1}\mathbf{B}\mathbf{C}^{-1} \\ \hline -\widetilde{\mathbf{A}}^{-1}\mathbf{B}^{\mathsf{T}}\mathbf{A}^{-1} & \widetilde{\mathbf{A}}^{-1} \end{array}\right], \tag{E.20}$$

where the tilde implies a Schur complement:

$$\widetilde{\mathbf{A}} = \mathbf{C} - \mathbf{B}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{B} \qquad \widetilde{\mathbf{C}} = \mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}^{\mathsf{T}} \tag{E.21}$$

$$\widetilde{\mathbf{A}}^{-1} = \mathbf{C}^{-1} + \mathbf{C}^{-1}\mathbf{B}^{\mathsf{T}}\widetilde{\mathbf{C}}^{-1}\mathbf{B}\mathbf{C}^{-1} \qquad \widetilde{\mathbf{C}}^{-1} = \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\widetilde{\mathbf{A}}^{-1}\mathbf{B}^{\mathsf{T}}\mathbf{A}^{-1}. \tag{E.22}$$

Thus

$$\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_+} = \left[\begin{array}{c|c} \widetilde{\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_0}^{-1}}^{-1} & -\widetilde{\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_0}^{-1}}^{-1}\boldsymbol{\Sigma}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}}^{-1}\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_0}^{-1} \\ \hline -\widetilde{\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_1}^{-1}}^{-1}\boldsymbol{\Sigma}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}}^{-\mathsf{T}}\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_1}^{-1} & \widetilde{\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_1}^{-1}}^{-1} \end{array}\right], \tag{E.23}$$

where $\widetilde{\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_0}^{-1}}^{-1} = \left(\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_1}^{-1} - \boldsymbol{\Sigma}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}}^{-1}\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_0}\boldsymbol{\Sigma}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}}^{-\mathsf{T}}\right)^{-1}$

Now from (8.8), the mean of the full posterior is

$$\boldsymbol{\mu}_{\hat{\mathbf{f}}_+} = \boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_+}\left(\mathbf{H}^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{g}\right) \tag{E.24}$$

$$= \left[\begin{array}{c} \widetilde{\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_0}^{-1}}^{-1}\mathbf{H}_1^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{g} - \widetilde{\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_0}^{-1}}^{-1}\boldsymbol{\Sigma}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}}^{-1}\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_0}\mathbf{H}_0^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{g} \\ \hline -\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_0}^{1}\boldsymbol{\Sigma}_{\widehat{\mathbf{f}_1,\mathbf{f}_0}}^{-\mathsf{T}}\widetilde{\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_0}^{-1}}^{-1}\mathbf{H}_1^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{g}^{-1}\mathbf{H}_0^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{g} \end{array}\right] \tag{E.25}$$

and the marginal distribution is just found using the first row of $\boldsymbol{\mu}_{\hat{\mathbf{f}}_+}$ and the first row and

column of $\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_+}$, i.e. $\mathrm{p}\left(\mathbf{f}_1 \,|\, \mathbf{g}\right) \propto \mathcal{N}\left(\mathbf{f}_1 \,\middle|\, \boldsymbol{\mu}_{\hat{\mathbf{f}}}, \boldsymbol{\Sigma}_{\widehat{\mathbf{f}}}\right)$ with

$$\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}}^{-1} = \widetilde{\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_0}^{-1}}^{-1} \tag{E.26}$$

$$\boldsymbol{\mu}_{\hat{\mathbf{f}}} = \widetilde{\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_0}^{-1}}^{-1} \mathbf{H}_1^{\mathsf{T}} \mathbf{Q}_w^{-1} \mathbf{g} - \widetilde{\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_0}^{-1}}^{-1} \boldsymbol{\Sigma}_{\mathbf{f}_1, \widehat{\mathbf{f}}_0}^{-1} \boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_0} \mathbf{H}_0^{\mathsf{T}} \mathbf{Q}_w^{-1} \mathbf{g}. \tag{E.27}$$

which is as in the previous section, other than we assumed $\bar{\mathbf{f}}_0$ was zero in the prior here.

E.3 Equivalent formulation

In this section, we show how the result for the marginal distribution may be reformulated to give the result obtained in another manner by Calvetti & Somersalo in [37]. The proof is surprisingly involved; essentially it requires several applications of the matrix inversion lemma and regrouping of terms. To simplify the derivation, we first find a modified version of the inversion formula in (E.20), letting $\mathbf{C} = -\mathbf{C}_2$ and $\widetilde{\mathbf{A}} = -\widetilde{\mathbf{A}_2}$:

$$\left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{B}^{\mathsf{T}} & -\mathbf{C}_2 \end{array}\right]^{-1} = \left[\begin{array}{c|c} \widetilde{\mathbf{C}}^{-1} & \widetilde{\mathbf{C}}^{-1}\mathbf{B}\mathbf{C}_2^{-1} \\ \hline \widetilde{\mathbf{A}_2}^{-1}\mathbf{B}^{\mathsf{T}}\mathbf{A}^{-1} & -\widetilde{\mathbf{A}_2}^{-1} \end{array}\right], \tag{E.28}$$

which results in:

$$\widetilde{\mathbf{A}_2} = \mathbf{C}_2 + \mathbf{B}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{B} \qquad \widetilde{\mathbf{C}} = \mathbf{A} + \mathbf{B}\mathbf{C}_2^{-1}\mathbf{B}^{\mathsf{T}} \tag{E.29}$$

$$\widetilde{\mathbf{A}_2}^{-1} = \mathbf{C}_2^{-1} - \mathbf{C}_2^{-1}\mathbf{B}^{\mathsf{T}}\widetilde{\mathbf{C}}^{-1}\mathbf{B}\mathbf{C}_2^{-1} \qquad \widetilde{\mathbf{C}}^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}\widetilde{\mathbf{A}_2}^{-1}\mathbf{B}^{\mathsf{T}}\mathbf{A}^{-1}. \tag{E.30}$$

which we can use to invert the following matrix:

$$\left[\begin{array}{c|c} \mathbf{Q}_w & \mathbf{H}_0 \\ \hline \mathbf{H}_0^{\mathsf{T}} & -\boldsymbol{\Sigma}_{\mathbf{f}_0}^{-1} \end{array}\right]^{-1} = \left[\begin{array}{c|c} \mathbf{Q}_{\widehat{w}}^{-1} & \mathbf{Q}_{\widehat{w}}^{-1}\mathbf{H}_0\boldsymbol{\Sigma}_{\mathbf{f}_0} \\ \hline (\mathbf{Q}_w^{-1}\mathbf{H}_0\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_0})^{\mathsf{T}} & -\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_0} \end{array}\right], \tag{E.31}$$

where we define the symbol $\mathbf{Q}_{\widehat{w}}$ as the Schur complement $\widetilde{\mathbf{C}}$ for this matrix, and the other Schur complement $\widetilde{\mathbf{A}_2}$ is equal to $\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_0}^{-1}$. We can expand these using (E.29) and (E.30) as:

$$\mathbf{Q}_{\widehat{w}} = \mathbf{Q}_w + \mathbf{H}_0\boldsymbol{\Sigma}_{\mathbf{f}_0}\mathbf{H}_0^{\mathsf{T}} = \left(\mathbf{Q}_w^{-1} - \mathbf{Q}_w^{-1}\mathbf{H}_0\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_0}^{-1}\mathbf{H}_0^{\mathsf{T}}\mathbf{Q}_w^{-1}\right)^{-1} \tag{E.32}$$

$$\boldsymbol{\Sigma}_{\widehat{\mathbf{f}}_0}^{-1} = \boldsymbol{\Sigma}_{\mathbf{f}_0}^{-1} + \mathbf{H}_0^{\mathsf{T}}\mathbf{Q}_w^{-1}\mathbf{H}_0 = \left(\boldsymbol{\Sigma}_{\mathbf{f}_0} - \boldsymbol{\Sigma}_{\mathbf{f}_0}\mathbf{H}_0^{\mathsf{T}}\mathbf{Q}_{\widehat{w}}^{-1}\mathbf{H}_0\boldsymbol{\Sigma}_{\mathbf{f}_0}\right)^{-1} \tag{E.33}$$

A further identity is produced by equating the upper right block and the transpose of the lower left (since the inverse matrix is also symmetric):

$$\mathbf{Q}_{\mathcal{W}}^{-1}\mathbf{H}_0\boldsymbol{\Sigma}_{\widehat{\mathrm{f}}_0} = \mathbf{Q}_{\widehat{\mathcal{W}}}^{-1}\mathbf{H}_0\boldsymbol{\Sigma}_{\mathrm{f}_0} \tag{E.34}$$

With these expressions, we can expand the covariance of $\mathrm{p}\left(\mathbf{f}_1 \mid \mathbf{g}\right)$ in (E.26) as follows:

$$\boldsymbol{\Sigma}_{\widehat{\mathrm{f}}} = \widetilde{\boldsymbol{\Sigma}_{\widehat{\mathrm{f}}_0}} = \left(\boldsymbol{\Sigma}_{\widehat{\mathrm{f}}_1}^{-1} - \boldsymbol{\Sigma}_{\widehat{\mathrm{f}_1,\mathrm{f}_0}}^{-1}\boldsymbol{\Sigma}_{\widehat{\mathrm{f}}_0}\boldsymbol{\Sigma}_{\widehat{\mathrm{f}_1,\mathrm{f}_0}}^{-\mathsf{T}}\right) \tag{E.35}$$

$$= \left(\boldsymbol{\Sigma}_{\mathrm{f}_1}^{-1} + \mathbf{H}_1^{\mathsf{T}}\mathbf{Q}_{\mathcal{W}}^{-1}\mathbf{H}_1)\right) - \left(\boldsymbol{\Sigma}_{\mathrm{f}_1,\mathrm{f}_0}^{-1} + \mathbf{H}_1^{\mathsf{T}}\mathbf{Q}_{\mathcal{W}}^{-1}\mathbf{H}_0\right)\boldsymbol{\Sigma}_{\widehat{\mathrm{f}}_0}\left(\boldsymbol{\Sigma}_{\mathrm{f}_1,\mathrm{f}_0}^{-1} + \mathbf{H}_1^{\mathsf{T}}\mathbf{Q}_{\mathcal{W}}^{-1}\mathbf{H}_0\right)^{\mathsf{T}} \tag{E.36}$$

$$\begin{aligned} &= \left(\boldsymbol{\Sigma}_{\mathrm{f}_1}^{-1} + \mathbf{H}_1^{\mathsf{T}}\mathbf{Q}_{\mathcal{W}}^{-1}\mathbf{H}_1)\right) - \boldsymbol{\Sigma}_{\mathrm{f}_1,\mathrm{f}_0}^{-1}\boldsymbol{\Sigma}_{\widehat{\mathrm{f}}_0}\boldsymbol{\Sigma}_{\mathrm{f}_1,\mathrm{f}_0}^{-\mathsf{T}} \\ &\quad - \mathbf{H}_1^{\mathsf{T}}\mathbf{Q}_{\mathcal{W}}^{-1}\mathbf{H}_0\boldsymbol{\Sigma}_{\widehat{\mathrm{f}}_0}\mathbf{H}_0^{\mathsf{T}}\mathbf{Q}_{\mathcal{W}}^{-1}\mathbf{H}_1 - \boldsymbol{\Sigma}_{\mathrm{f}_1,\mathrm{f}_0}^{-1}\boldsymbol{\Sigma}_{\widehat{\mathrm{f}}_0}\mathbf{H}_0^{\mathsf{T}}\mathbf{Q}_{\mathcal{W}}^{-1}\mathbf{H}_1 - \mathbf{H}_1^{\mathsf{T}}\mathbf{Q}_{\mathcal{W}}^{-1}\mathbf{H}_0\boldsymbol{\Sigma}_{\widehat{\mathrm{f}}_0}\boldsymbol{\Sigma}_{\mathrm{f}_1,\mathrm{f}_0}^{-\mathsf{T}} \end{aligned} \tag{E.37}$$

$$\begin{aligned} &= \boldsymbol{\Sigma}_{\mathrm{f}_1}^{-1} - \boldsymbol{\Sigma}_{\mathrm{f}_1,\mathrm{f}_0}^{-1}\left(\boldsymbol{\Sigma}_{\mathrm{f}_0} - \boldsymbol{\Sigma}_{\mathrm{f}_0}\mathbf{H}_0^{\mathsf{T}}\mathbf{Q}_{\widehat{\mathcal{W}}}^{-1}\mathbf{H}_0\boldsymbol{\Sigma}_{\mathrm{f}_0}\right)\boldsymbol{\Sigma}_{\mathrm{f}_1,\mathrm{f}_0}^{-\mathsf{T}} + \mathbf{H}_1^{\mathsf{T}}\left(\mathbf{Q}_{\mathcal{W}}^{-1} - \mathbf{Q}_{\mathcal{W}}^{-1}\mathbf{H}_0\boldsymbol{\Sigma}_{\widehat{\mathrm{f}}_0}\mathbf{H}_0^{\mathsf{T}}\mathbf{Q}_{\mathcal{W}}^{-1}\right)\mathbf{H}_1 \\ &\quad - \boldsymbol{\Sigma}_{\mathrm{f}_1,\mathrm{f}_0}^{-1}\boldsymbol{\Sigma}_{\mathrm{f}_0}\mathbf{H}_0^{\mathsf{T}}\mathbf{Q}_{\widehat{\mathcal{W}}}^{-1}\mathbf{H}_1 - \mathbf{H}_1^{\mathsf{T}}\mathbf{Q}_{\widehat{\mathcal{W}}}^{-1}\mathbf{H}_0\boldsymbol{\Sigma}_{\mathrm{f}_0}\boldsymbol{\Sigma}_{\mathrm{f}_1,\mathrm{f}_0}^{-\mathsf{T}} \end{aligned} \tag{E.38}$$

where we have substituted for $\boldsymbol{\Sigma}_{\widehat{\mathrm{f}}_0}$ using (E.30) in the second term, and the last two terms have made use of the equivalence in (E.34). The portion of the third term within parentheses is just equal to $\mathbf{Q}_{\widehat{\mathcal{W}}}^{-1}$. Hence we have:

$$\begin{aligned} \boldsymbol{\Sigma}_{\widehat{\mathrm{f}}} &= \left(\boldsymbol{\Sigma}_{\mathrm{f}_1}^{-1} - \boldsymbol{\Sigma}_{\mathrm{f}_1,\mathrm{f}_0}^{-1}\boldsymbol{\Sigma}_{\mathrm{f}_0}\boldsymbol{\Sigma}_{\mathrm{f}_1,\mathrm{f}_0}^{-\mathsf{T}}\right) + \boldsymbol{\Sigma}_{\mathrm{f}_1,\mathrm{f}_0}^{-1}\boldsymbol{\Sigma}_{\mathrm{f}_0}\mathbf{H}_0^{\mathsf{T}}\mathbf{Q}_{\widehat{\mathcal{W}}}^{-1}\mathbf{H}_0\boldsymbol{\Sigma}_{\mathrm{f}_0}\boldsymbol{\Sigma}_{\mathrm{f}_1,\mathrm{f}_0}^{-\mathsf{T}} + \mathbf{H}_1^{\mathsf{T}}\mathbf{Q}_{\widehat{\mathcal{W}}}^{-1}\mathbf{H}_1 \\ &\quad - \boldsymbol{\Sigma}_{\mathrm{f}_1,\mathrm{f}_0}^{-1}\boldsymbol{\Sigma}_{\mathrm{f}_0}\mathbf{H}_0^{\mathsf{T}}\mathbf{Q}_{\widehat{\mathcal{W}}}^{-1}\mathbf{H}_1 - \mathbf{H}_1^{\mathsf{T}}\mathbf{Q}_{\widehat{\mathcal{W}}}^{-1}\mathbf{H}_0\boldsymbol{\Sigma}_{\mathrm{f}_0}\boldsymbol{\Sigma}_{\mathrm{f}_1,\mathrm{f}_0}^{-\mathsf{T}} \end{aligned} \tag{E.39}$$

$$= \widetilde{\boldsymbol{\Sigma}_{\mathrm{f}_0}^{-1}} + \left(\mathbf{H}_1^{\mathsf{T}} - \boldsymbol{\Sigma}_{\mathrm{f}_1,\mathrm{f}_0}^{-1}\boldsymbol{\Sigma}_{\mathrm{f}_0}\mathbf{H}_0^{\mathsf{T}}\right)\mathbf{Q}_{\widehat{\mathcal{W}}}^{-1}\left(\mathbf{H}_1 - \mathbf{H}_0\boldsymbol{\Sigma}_{\mathrm{f}_0}\boldsymbol{\Sigma}_{\mathrm{f}_1,\mathrm{f}_0}^{-\mathsf{T}}\right) \tag{E.40}$$

$$\tag{E.41}$$

where $\widetilde{\boldsymbol{\Sigma}_{\mathrm{f}_0}^{-1}}$ is the Schur complement of $\boldsymbol{\Sigma}_{\mathrm{f}_0}$ in the matrix $\boldsymbol{\Sigma}_{\mathrm{f}_+}$, i.e. it is actually the partition of the prior covariance for $\mathbf{f}$ corresponding to the region $\mathcal{R}$.

We can simplify the notation by denoting the partitioning of the covariance $\boldsymbol{\Sigma}_{\mathrm{f}_+}$ as

$$\boldsymbol{\Sigma}_{\mathrm{f}_+} = \left[\begin{array}{c|c} \boldsymbol{\Gamma}_{\mathrm{f}_1} & \boldsymbol{\Gamma}_{\mathrm{f}_1,\mathrm{f}_0} \\ \hline \boldsymbol{\Gamma}_{\mathrm{f}_1,\mathrm{f}_0}^{\mathsf{T}} & \boldsymbol{\Gamma}_{\mathrm{f}_0} \end{array}\right] = \left[\begin{array}{c|c} \widetilde{\boldsymbol{\Sigma}_{\mathrm{f}_0}^{-1}}^{-1} & -\widetilde{\boldsymbol{\Sigma}_{\mathrm{f}_0}^{-1}}^{-1}\boldsymbol{\Sigma}_{\mathrm{f}_1,\mathrm{f}_0}^{-1}\boldsymbol{\Sigma}_{\mathrm{f}_0} \\ \hline -\widetilde{\boldsymbol{\Sigma}_{\mathrm{f}_1}^{-1}}^{-1}\boldsymbol{\Sigma}_{\mathrm{f}_1,\mathrm{f}_0}^{-\mathsf{T}}\boldsymbol{\Sigma}_{\mathrm{f}_1} & \widetilde{\boldsymbol{\Sigma}_{\mathrm{f}_1}^{-1}}^{-1} \end{array}\right] \tag{E.42}$$

$$\Rightarrow \boldsymbol{\Sigma}_{\mathrm{f}_+}^{-1} = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{\mathrm{f}_1}^{-1} & \boldsymbol{\Sigma}_{\mathrm{f}_1,\mathrm{f}_0}^{-1} \\ \hline \boldsymbol{\Sigma}_{\mathrm{f}_1,\mathrm{f}_0}^{-\mathsf{T}} & \boldsymbol{\Sigma}_{\mathrm{f}_0}^{-1} \end{array}\right] = \left[\begin{array}{c|c} \widetilde{\boldsymbol{\Gamma}_{\mathrm{f}_0}}^{-1} & -\widetilde{\boldsymbol{\Gamma}_{\mathrm{f}_0}}\boldsymbol{\Gamma}_{\mathrm{f}_1,\mathrm{f}_0}\boldsymbol{\Gamma}_{\mathrm{f}_0}^{-1} \\ \hline -\widetilde{\boldsymbol{\Gamma}_{\mathrm{f}_1}}^{-1}\boldsymbol{\Gamma}_{\mathrm{f}_1,\mathrm{f}_0}^{\mathsf{T}}\boldsymbol{\Gamma}_{\mathrm{f}_1}^{-1} & \widetilde{\boldsymbol{\Gamma}_{\mathrm{f}_1}}^{-1} \end{array}\right]. \tag{E.43}$$

This then gives

$$\begin{aligned} \boldsymbol{\Sigma}_{\hat{\mathrm{f}}} = \widetilde{\boldsymbol{\Sigma}_{\mathrm{f}_0}^{-1}} + \widetilde{\boldsymbol{\Sigma}_{\mathrm{f}_0}^{-1}} \left(\widetilde{\boldsymbol{\Sigma}_{\mathrm{f}_0}^{-1}}^{-1}\mathbf{H}_1^{\mathsf{T}} - \widetilde{\boldsymbol{\Sigma}_{\mathrm{f}_0}^{-1}}^{-1}\boldsymbol{\Sigma}_{\mathrm{f}_1,\mathrm{f}_0}^{-1}\boldsymbol{\Sigma}_{\mathrm{f}_0}\mathbf{H}_0^{\mathsf{T}}\right)\mathbf{Q}_{\widehat{\mathcal{W}}}^{-1}\cdot \\ \cdot\left(\widetilde{\boldsymbol{\Sigma}_{\mathrm{f}_0}^{-1}}^{-1}\mathbf{H}_1^{\mathsf{T}} - \widetilde{\boldsymbol{\Sigma}_{\mathrm{f}_0}^{-1}}^{-1}\boldsymbol{\Sigma}_{\mathrm{f}_1,\mathrm{f}_0}^{-1}\boldsymbol{\Sigma}_{\mathrm{f}_0}\mathbf{H}_0^{\mathsf{T}}\right)^{-\mathsf{T}}\widetilde{\boldsymbol{\Sigma}_{\mathrm{f}_0}^{-1}} \end{aligned} \tag{E.44}$$

$$= \boldsymbol{\Gamma}_{\mathrm{f}_1}^{-1} + \boldsymbol{\Gamma}_{\mathrm{f}_1}^{-1}\left(\boldsymbol{\Gamma}_{\mathrm{f}_1}\mathbf{H}_1^{\mathsf{T}} + \boldsymbol{\Gamma}_{\mathrm{f}_1,\mathrm{f}_0}\mathbf{H}_0^{\mathsf{T}}\right)\mathbf{Q}_{\widehat{\mathcal{W}}}^{-1}\left(\boldsymbol{\Gamma}_{\mathrm{f}_1}\mathbf{H}_1^{\mathsf{T}} + \boldsymbol{\Gamma}_{\mathrm{f}_1,\mathrm{f}_0}\mathbf{H}_0^{\mathsf{T}}\right)^{\mathsf{T}}\boldsymbol{\Gamma}_{\mathrm{f}_1}^{-1} \tag{E.45}$$

$$= \boldsymbol{\Gamma}_{\mathrm{f}_1}^{-1} + \boldsymbol{\Gamma}_{\mathrm{f}_1}^{-1}\boldsymbol{\Gamma}_{\mathrm{H}}^{\mathsf{T}}\mathbf{Q}_{\widehat{\mathcal{W}}}^{-1}\boldsymbol{\Gamma}_{\mathrm{H}}\boldsymbol{\Gamma}_{\mathrm{f}_1}^{-1}, \tag{E.46}$$

where we define $\boldsymbol{\Gamma}_{\mathrm{H}} = \left(\boldsymbol{\Gamma}_{\mathrm{f}_1}\mathbf{H}_1^{\mathsf{T}} + \boldsymbol{\Gamma}_{\mathrm{f}_1,\mathrm{f}_0}\mathbf{H}_0^{\mathsf{T}}\right)^{\mathsf{T}}$.

Now we can rewrite (E.27) as follows:

$$\boldsymbol{\mu}_{\hat{\mathrm{f}}} = \boldsymbol{\Sigma}_{\hat{\mathrm{f}}}^{-1}\left(\mathbf{H}_1^{\mathsf{T}}\mathbf{Q}_{\mathcal{W}}^{-1} - \boldsymbol{\Sigma}_{\widehat{\mathrm{f}_1,\mathrm{f}_0}}^{-1}\boldsymbol{\Sigma}_{\hat{\mathrm{f}}_0}\mathbf{H}_0^{\mathsf{T}}\mathbf{Q}_{\mathcal{W}}^{-1}\right)\mathbf{g}, \tag{E.47}$$

and the term in brackets is

$$\mathbf{H}_1^{\mathsf{T}}\mathbf{Q}_{\mathcal{W}}^{-1} - \left(\boldsymbol{\Sigma}_{\mathrm{f}_1,\mathrm{f}_0}^{-1} + \mathbf{H}_1^{\mathsf{T}}\mathbf{Q}_{\mathcal{W}}^{-1}\mathbf{H}_0\right)\boldsymbol{\Sigma}_{\hat{\mathrm{f}}_0}\mathbf{H}_0^{\mathsf{T}}\mathbf{Q}_{\mathcal{W}}^{-1} \tag{E.48}$$

$$= \mathbf{H}_1^{\mathsf{T}}\left(\mathbf{Q}_{\mathcal{W}}^{-1} - \mathbf{Q}_{\mathcal{W}}^{-1}\mathbf{H}_0\boldsymbol{\Sigma}_{\hat{\mathrm{f}}_0}\mathbf{H}_0^{\mathsf{T}}\mathbf{Q}_{\mathcal{W}}^{-1}\right) - \boldsymbol{\Sigma}_{\mathrm{f}_1,\mathrm{f}_0}^{-1}\boldsymbol{\Sigma}_{\hat{\mathrm{f}}_0}\mathbf{H}_0^{\mathsf{T}}\mathbf{Q}_{\mathcal{W}}^{-1} \tag{E.49}$$

$$= \mathbf{H}_1^{\mathsf{T}}\mathbf{Q}_{\widehat{\mathcal{W}}}^{-1} - \boldsymbol{\Sigma}_{\mathrm{f}_1,\mathrm{f}_0}^{-1}\boldsymbol{\Sigma}_{\mathrm{f}_0}\mathbf{H}_0^{\mathsf{T}}\mathbf{Q}_{\widehat{\mathcal{W}}}^{-1} \tag{E.50}$$

$$= \left(\mathbf{H}_1^{\mathsf{T}} - \boldsymbol{\Sigma}_{\mathrm{f}_1,\mathrm{f}_0}^{-1}\boldsymbol{\Sigma}_{\mathrm{f}_0}\mathbf{H}_0^{\mathsf{T}}\right)\mathbf{Q}_{\widehat{\mathcal{W}}}^{-1} \tag{E.51}$$

$$= \widetilde{\boldsymbol{\Sigma}_{\mathrm{f}_0}^{-1}}\left(\widetilde{\boldsymbol{\Sigma}_{\mathrm{f}_0}^{-1}}^{-1}\mathbf{H}_1^{\mathsf{T}} - \widetilde{\boldsymbol{\Sigma}_{\mathrm{f}_0}^{-1}}^{-1}\boldsymbol{\Sigma}_{\mathrm{f}_1,\mathrm{f}_0}^{-1}\boldsymbol{\Sigma}_{\mathrm{f}_0}\mathbf{H}_0^{\mathsf{T}}\right)\mathbf{Q}_{\widehat{\mathcal{W}}}^{-1} \tag{E.52}$$

$$= \boldsymbol{\Gamma}_{\mathrm{f}_1}^{-1}\left(\boldsymbol{\Gamma}_{\mathrm{f}_1}\mathbf{H}_1^{\mathsf{T}} + \boldsymbol{\Gamma}_{\mathrm{f}_1,\mathrm{f}_0}\mathbf{H}_0^{\mathsf{T}}\right)\mathbf{Q}_{\widehat{\mathcal{W}}}^{-1} \tag{E.53}$$

$$= \left(\mathbf{Q}_{\widehat{\mathcal{W}}}(\boldsymbol{\Gamma}_{\mathrm{f}_1}\mathbf{H}_1^{\mathsf{T}} + \boldsymbol{\Gamma}_{\mathrm{f}_1,\mathrm{f}_0}\mathbf{H}_0^{\mathsf{T}})^{-1}\boldsymbol{\Gamma}_{\mathrm{f}_1}\right)^{-1} \tag{E.54}$$

$$= \left(\mathbf{Q}_{\widehat{\mathcal{W}}}\boldsymbol{\Gamma}_{\mathrm{H}}^{-\mathsf{T}}\boldsymbol{\Gamma}_{\mathrm{f}_1}\right)^{-1} \tag{E.55}$$

therefore, by substituting (E.46), then applying the identity $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$ twice and rearranging,

$$\boldsymbol{\mu}_{\hat{\mathrm{f}}} = \left(\boldsymbol{\Gamma}_{\mathrm{f}_1}^{-1} + \boldsymbol{\Gamma}_{\mathrm{f}_1}^{-1}\boldsymbol{\Gamma}_{\mathrm{H}}^{\mathsf{T}}\mathbf{Q}_{\widehat{\mathcal{W}}}^{-1}\boldsymbol{\Gamma}_{\mathrm{H}}\boldsymbol{\Gamma}_{\mathrm{f}_1}^{-1}\right)^{-1}\left(\mathbf{Q}_{\widehat{\mathcal{W}}}\boldsymbol{\Gamma}_{\mathrm{H}}^{-\mathsf{T}}\boldsymbol{\Gamma}_{\mathrm{f}_1}\right)^{-1}\mathbf{g} \tag{E.56}$$

$$= \left(\left(\mathbf{Q}_{\widehat{w}}\boldsymbol{\Gamma}_{\mathrm{H}}^{-\mathrm{T}}\boldsymbol{\Gamma}_{\mathrm{f}_1}\right)\left(\boldsymbol{\Gamma}_{\mathrm{f}_1}^{-1} + \boldsymbol{\Gamma}_{\mathrm{f}_1}^{-1}\boldsymbol{\Gamma}_{\mathrm{H}}^{\mathrm{T}}\mathbf{Q}_{\widehat{w}}^{-1}\boldsymbol{\Gamma}_{\mathrm{H}}\boldsymbol{\Gamma}_{\mathrm{f}_1}^{-1}\right)\right)^{-1}\mathbf{g} \tag{E.57}$$

$$= \left(\mathbf{Q}_{\widehat{w}}\boldsymbol{\Gamma}_{\mathrm{H}}^{-\mathrm{T}} + \boldsymbol{\Gamma}_{\mathrm{H}}\boldsymbol{\Gamma}_{\mathrm{f}_1}^{-1}\right)^{-1}\mathbf{g} \tag{E.58}$$

$$= \left(\left(\mathbf{Q}_{\widehat{w}} + \boldsymbol{\Gamma}_{\mathrm{H}}\boldsymbol{\Gamma}_{\mathrm{f}_1}^{-1}\boldsymbol{\Gamma}_{\mathrm{H}}^{\mathrm{T}}\right)\boldsymbol{\Gamma}_{\mathrm{H}}^{-\mathrm{T}}\right)^{-1}\mathbf{g} \tag{E.59}$$

$$= \boldsymbol{\Gamma}_{\mathrm{H}}^{\mathrm{T}}\left(\mathbf{Q}_{\widehat{w}} + \boldsymbol{\Gamma}_{\mathrm{H}}\boldsymbol{\Gamma}_{\mathrm{f}_1}^{-1}\boldsymbol{\Gamma}_{\mathrm{H}}^{\mathrm{T}}\right)^{-1}\mathbf{g} \tag{E.60}$$

$$= \boldsymbol{\Gamma}_{\mathrm{H}}^{\mathrm{T}}\left(\mathbf{Q}_{w} + \mathbf{H}_0\boldsymbol{\Sigma}_{\mathrm{f}_0}\mathbf{H}_0^{\mathrm{T}} + \boldsymbol{\Gamma}_{\mathrm{H}}\boldsymbol{\Gamma}_{\mathrm{f}_1}^{-1}\boldsymbol{\Gamma}_{\mathrm{H}}^{\mathrm{T}}\right)^{-1} \tag{E.61}$$

Then substituting $\boldsymbol{\Sigma}_{\mathrm{f}_0} = \widetilde{\boldsymbol{\Gamma}_{\mathrm{f}_1}} = \left(\boldsymbol{\Gamma}_{\mathrm{f}_0} - \boldsymbol{\Gamma}_{\mathrm{f}_1,\mathrm{f}_0}^{\mathrm{T}}\boldsymbol{\Gamma}_{\mathrm{f}_1}^{-1}\boldsymbol{\Gamma}_{\mathrm{f}_1,\mathrm{f}_0}\right)$ and $\boldsymbol{\Gamma}_{\mathrm{H}} = \left(\mathbf{H}_1\boldsymbol{\Gamma}_{\mathrm{f}_1} - \mathbf{H}_0\boldsymbol{\Gamma}_{\mathrm{f}_1,\mathrm{f}_0}^{\mathrm{T}}\right)$, gives

$$\begin{aligned}\boldsymbol{\mu}_{\hat{\mathrm{f}}} = \boldsymbol{\Gamma}_{\mathrm{H}}^{\mathrm{T}}\Bigg(&\mathbf{Q}_{w} + \mathbf{H}_0\left(\boldsymbol{\Gamma}_{\mathrm{f}_0} - \boldsymbol{\Gamma}_{\mathrm{f}_1,\mathrm{f}_0}^{\mathrm{T}}\boldsymbol{\Gamma}_{\mathrm{f}_1}^{-1}\boldsymbol{\Gamma}_{\mathrm{f}_1,\mathrm{f}_0}\right)\mathbf{H}_0^{\mathrm{T}} \\ &+ \left(\mathbf{H}_1\boldsymbol{\Gamma}_{\mathrm{f}_1} + \mathbf{H}_0\boldsymbol{\Gamma}_{\mathrm{f}_1,\mathrm{f}_0}^{\mathrm{T}}\right)\boldsymbol{\Gamma}_{\mathrm{f}_1}^{-1}\left(\mathbf{H}_1\boldsymbol{\Gamma}_{\mathrm{f}_1} + \mathbf{H}_0\boldsymbol{\Gamma}_{\mathrm{f}_1,\mathrm{f}_0}^{\mathrm{T}}\right)^{\mathrm{T}}\Bigg)^{-1}\mathbf{g}\end{aligned} \tag{E.62}$$

$$= \boldsymbol{\Gamma}_{\mathrm{H}}^{\mathrm{T}}\left(\mathbf{Q}_{w} + \mathbf{H}_0\boldsymbol{\Gamma}_{\mathrm{f}_0}\mathbf{H}_0^{\mathrm{T}} + \mathbf{H}_1\boldsymbol{\Gamma}_{\mathrm{f}_1}\mathbf{H}_1^{\mathrm{T}} + \mathbf{H}_1\boldsymbol{\Gamma}_{\mathrm{f}_1,\mathrm{f}_0}\mathbf{H}_0^{\mathrm{T}} + \mathbf{H}_0\boldsymbol{\Gamma}_{\mathrm{f}_1,\mathrm{f}_0}^{\mathrm{T}}\mathbf{H}_1^{\mathrm{T}}\right)^{-1}\mathbf{g} \tag{E.63}$$

$$= \left(\boldsymbol{\Gamma}_{\mathrm{f}_1}\mathbf{H}_1^{\mathrm{T}} + \boldsymbol{\Gamma}_{\mathrm{f}_1,\mathrm{f}_0}\mathbf{H}_0^{\mathrm{T}}\right)\left(\mathbf{Q}_{w} + \mathbf{H}\boldsymbol{\Sigma}_{\mathrm{f}_+}\mathbf{H}^{\mathrm{T}}\right)^{-1}\mathbf{g} \tag{E.64}$$

This last result is equivalent to equation 14 in [37].

Finally note that with the above definitions, this final expression is also derived directly; by first inverting $\boldsymbol{\Sigma}_{\widehat{\mathrm{f}}_+}^{-1} = \boldsymbol{\Sigma}_{\mathrm{f}_+}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{Q}_{w}^{-1}\mathbf{H}$ with the matrix inversion lemma to give

$$\boldsymbol{\Sigma}_{\widehat{\mathrm{f}}_+} = \boldsymbol{\Sigma}_{\mathrm{f}_+} - \boldsymbol{\Sigma}_{\mathrm{f}_+}\mathbf{H}^{\mathrm{T}}\left(\mathbf{Q}_{w} + \mathbf{H}\boldsymbol{\Sigma}_{\mathrm{f}_+}\mathbf{H}^{\mathrm{T}}\right)^{-1}\mathbf{H}\boldsymbol{\Sigma}_{\mathrm{f}_+} \tag{E.65}$$

then by considering the top block of

$$\boldsymbol{\mu}_{\hat{\mathrm{f}}_+} = \boldsymbol{\Sigma}_{\widehat{\mathrm{f}}_+}\left(\mathbf{H}^{\mathrm{T}}\mathbf{Q}_{w}^{-1}\mathbf{g}\right) \tag{E.66}$$

$$= \left(\boldsymbol{\Sigma}_{\mathrm{f}_+} - \boldsymbol{\Sigma}_{\mathrm{f}_+}\mathbf{H}^{\mathrm{T}}\left(\mathbf{Q}_{w} + \mathbf{H}\boldsymbol{\Sigma}_{\mathrm{f}_+}\mathbf{H}^{\mathrm{T}}\right)^{-1}\mathbf{H}\boldsymbol{\Sigma}_{\mathrm{f}_+}\right)\mathbf{H}^{\mathrm{T}}\mathbf{Q}_{w}^{-1}\mathbf{g} \tag{E.67}$$

$$= \boldsymbol{\Sigma}_{\mathrm{f}_+}\mathbf{H}^{\mathrm{T}}\left(\mathbf{Q}_{w}^{-1} - \left(\mathbf{Q}_{w} + \mathbf{H}\boldsymbol{\Sigma}_{\mathrm{f}_+}\mathbf{H}^{\mathrm{T}}\right)^{-1}\mathbf{H}\boldsymbol{\Sigma}_{\mathrm{f}_+}\mathbf{H}^{\mathrm{T}}\mathbf{Q}_{w}^{-1}\right)\mathbf{g} \tag{E.68}$$

$$= \boldsymbol{\Sigma}_{\mathrm{f}_+}\mathbf{H}^{\mathrm{T}}\left(\mathbf{Q}_{w}^{-1} - \left(\boldsymbol{\Sigma}_{\mathrm{f}_+}^{-1}\mathbf{H}^{-1}\mathbf{Q}_{w} + \mathbf{H}^{\mathrm{T}}\right)\mathbf{H}^{\mathrm{T}}\mathbf{Q}_{w}^{-1}\right)\mathbf{g} \tag{E.69}$$

$$= \boldsymbol{\Sigma}_{\mathrm{f}_+}\mathbf{H}^{\mathrm{T}}\left(\mathbf{Q}_{w}^{-1} - \mathbf{Q}_{w}^{-1}\mathbf{H}\left(\boldsymbol{\Sigma}_{\mathrm{f}_+}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{Q}_{w}^{-1}\mathbf{H}\right)\mathbf{H}^{\mathrm{T}}\mathbf{Q}_{w}^{-1}\right)\mathbf{g} \tag{E.70}$$

$$= \boldsymbol{\Sigma}_{\mathrm{f}_+}\mathbf{H}^{\mathrm{T}}\left(\mathbf{Q}_{w} + \mathbf{H}\boldsymbol{\Sigma}_{\mathrm{f}_+}\mathbf{H}^{\mathrm{T}}\right)^{-1}\mathbf{g}, \tag{E.71}$$

gives the desired result.

Appendix F

VB distribution approximations

Here we show the derivation of some of the approximate distributions for the proposed model, used under the VB estimation method.

F.1 Distribution Approximation for $\mathbf{f}$

For the distribution $\mathrm{q}^{\mathrm{k}}(\mathbf{f})$, using (5.25) (ignoring terms independent of $\mathbf{f}$), we find:

$$\log \mathrm{q}^{\mathrm{k}}(\mathbf{f}) \propto \mathbb{E}\left[\log\left(\mathrm{p}(\mathbf{g}\,|\,\boldsymbol{\Theta})\,\mathrm{p}(\mathbf{f}\,|\,\boldsymbol{\alpha}_{\mathrm{f}}, \boldsymbol{\mu}, \sigma_{\mathrm{v}})\right)\right]_{\mathrm{q}^{\mathrm{k}}(\boldsymbol{\Theta}_{\setminus \mathbf{f}})} \tag{F.1}$$

$$\propto -\frac{1}{2}\left(\mathbb{E}\left[\sigma_{w}^{-2}\|\mathbf{g}-\mathbf{H}\mathbf{f}\|^{2}\right]_{\mathrm{q}^{\mathrm{k}}(\mathbf{h},\sigma_{w})} + \mathbb{E}\left[\|\boldsymbol{\nu}\|^{2}_{\mathbf{Q}_{\nu}}\right]_{\mathrm{q}^{\mathrm{k}}(\mathbf{m}_{\mathbf{f}},\boldsymbol{\alpha}_{\mathrm{f}},\sigma_{\mathrm{v}})}\right) \tag{F.2}$$

The first expectation is calculated as

$$\mathbb{E}\left[\sigma_{w}^{-2}\|\mathbf{g}-\mathbf{H}\mathbf{f}\|^{2}\right]_{\mathrm{q}^{\mathrm{k}}(\mathbf{h},\sigma_{w})} = \mathbb{E}^{\mathrm{k}}\left[\sigma_{w}^{-2}\right]\left(\mathbf{g}^{\mathsf{T}}\mathbf{g} - 2\mathbf{g}^{\mathsf{T}}\mathbb{E}^{\mathrm{k}}[\mathbf{H}]\,\mathbf{f} + \mathbf{f}^{\mathsf{T}}\mathbb{E}^{\mathrm{k}}\left[\mathbf{H}^{\mathsf{T}}\mathbf{H}\right]\mathbf{f}\right) \tag{F.3}$$

$$\begin{aligned} = \mathbb{E}^{\mathrm{k}}\left[\sigma_{w}^{-2}\right]\Big(&\mathbf{g}^{\mathsf{T}}\mathbf{g} - 2\mathbf{g}^{\mathsf{T}}\mathbb{E}^{\mathrm{k}}[\mathbf{H}]\,\mathbf{f} + \mathbf{f}^{\mathsf{T}}\mathbb{E}^{\mathrm{k}}[\mathbf{H}]^{\mathsf{T}}\,\mathbb{E}^{\mathrm{k}}[\mathbf{H}]\,\mathbf{f} \\ &+ \mathbf{f}^{\mathsf{T}}\Big(\mathbb{E}^{\mathrm{k}}\left[\mathbf{H}^{\mathsf{T}}\mathbf{H}\right] - \mathbb{E}^{\mathrm{k}}[\mathbf{H}]^{\mathsf{T}}\,\mathbb{E}^{\mathrm{k}}[\mathbf{H}]\Big)\mathbf{f}\Big) \end{aligned} \tag{F.4}$$

$$= \mathbb{E}^{\mathrm{k}}\left[\sigma_{w}^{-2}\right]\left(\|\mathbf{g} - \mathbb{E}^{\mathrm{k}}[\mathbf{H}]\,\mathbf{f}\|^{2} + \mathbf{f}^{\mathsf{T}}\,\mathrm{cov}^{\mathrm{k}}\left[\mathbf{H}^{\mathsf{T}}\right]\mathbf{f}\right), \tag{F.5}$$

where we have used the notation $\mathrm{cov}^{\mathrm{k}}\left[\mathbf{X}^{\mathsf{T}}\right] \triangleq \mathbb{E}^{\mathrm{k}}\left[\mathbf{X}^{\mathsf{T}}\mathbf{X} - \mathbb{E}^{\mathrm{k}}[\mathbf{X}]^{\mathsf{T}}\,\mathbb{E}^{\mathrm{k}}[\mathbf{X}]\right]$. Observe that while $\mathrm{cov}^{\mathrm{k}}[\mathbf{h}]$ is of dimension $\dim(\mathbf{h}) \times \dim(\mathbf{h})$, $\mathrm{cov}^{\mathrm{k}}\left[\mathbf{H}^{\mathsf{T}}\right]$ is of dimension $\dim(\mathbf{f}) \times \dim(\mathbf{f})$. See Appendix G for further details on this notation.

Considering the second expectation in (F.2), this may be expanded as

$$\mathbb{E}\left[\|\boldsymbol{\nu}\|^{2}_{\mathbf{Q}_{\nu}}\right]_{\mathrm{q}^{\mathrm{k}}(\mathbf{m}_{\mathbf{f}},\boldsymbol{\alpha}_{\mathrm{f}},\sigma_{\mathrm{v}})} = \mathbb{E}\left[\|\mathbf{f} - (\mathbf{I}-\mathbf{A}_{\mathrm{f}})^{-1}(\mathbf{I}-\mathbf{A}_{\mathrm{m}})\mathbf{m}_{\mathbf{f}}\|^{2}_{\boldsymbol{\Sigma}_{\mathrm{f}}}\right]_{\mathrm{q}^{\mathrm{k}}(\mathbf{m}_{\mathbf{f}},\boldsymbol{\alpha}_{\mathrm{f}},\sigma_{\mathrm{v}})} \tag{F.6}$$

$$= \mathbb{E}\left[\mathbf{f}^\mathsf{T}(\mathbf{I}-\mathbf{A}_\mathrm{f})^\mathsf{T}\mathbf{Q}_v^{-1}(\mathbf{I}-\mathbf{A}_\mathrm{f})\mathbf{f}\right]_{\mathrm{q}^\mathrm{k}(\boldsymbol{\alpha}_\mathrm{f},\boldsymbol{\sigma}_v)}$$
$$- 2\mathbb{E}\left[\mathbf{f}^\mathsf{T}(\mathbf{I}-\mathbf{A}_\mathrm{f})^\mathsf{T}\mathbf{Q}_v^{-1}(\mathbf{I}-\mathbf{A}_\mathrm{m})\mathbf{m}_\mathbf{f}\right]_{\mathrm{q}^\mathrm{k}(\mathbf{m}_\mathbf{f},\boldsymbol{\alpha}_\mathrm{f},\boldsymbol{\sigma}_v)} + \text{term independent of } \mathbf{f}. \tag{F.7}$$

The first term of (F.7) is found as

$$\mathbf{f}^\mathsf{T}\left(\mathbb{E}^\mathrm{k}\left[\mathbf{Q}_v^{-1}\right] - 2\mathbb{E}^\mathrm{k}\left[\mathbf{A}_\mathrm{f}\right]^\mathsf{T}\mathbb{E}^\mathrm{k}\left[\mathbf{Q}_v^{-1}\right] + \mathbb{E}\left[\mathbf{A}_\mathrm{f}^\mathsf{T}\mathbb{E}^\mathrm{k}\left[\mathbf{Q}_v^{-1}\right]\mathbf{A}_\mathrm{f}\right]_{\mathrm{q}^\mathrm{k}(\boldsymbol{\alpha}_\mathrm{f})}\right)\mathbf{f} \tag{F.8}$$
$$=\mathbf{f}^\mathsf{T}\left(\|\mathbf{I}-\mathbb{E}^\mathrm{k}\left[\mathbf{A}_\mathrm{f}\right]\|^2_{\mathbf{Q}_v} + \mathrm{cov}^\mathrm{k}\left[\mathbf{A}_\mathrm{f}^\mathsf{T}\mathbf{L}_v^\mathsf{T}\right]\right)\mathbf{f}, \tag{F.9}$$

where we have defined $\mathbf{L}_v^\mathsf{T}\mathbf{L}_v = \mathbb{E}^\mathrm{k}\left[\mathbf{Q}_v^{-1}\right]$. Similarly, the second term of (F.7) is

$$-2\mathbf{f}^\mathsf{T}\left(\mathbb{E}^\mathrm{k}\left[\mathbf{Q}_v^{-1}\right] - (\mathbb{E}^\mathrm{k}\left[\mathbf{A}_\mathrm{f}\right]^\mathsf{T} + \mathbb{E}^\mathrm{k}\left[\mathbf{A}_\mathrm{m}\right]^\mathsf{T})\mathbb{E}^\mathrm{k}\left[\mathbf{Q}_v^{-1}\right] + \mathbb{E}\left[\mathbf{A}_\mathrm{f}^\mathsf{T}\mathbb{E}^\mathrm{k}\left[\mathbf{Q}_v^{-1}\right]\mathbf{A}_\mathrm{m}\right]_{\mathrm{q}^\mathrm{k}(\boldsymbol{\alpha}_\mathrm{f})}\right)\mathbb{E}^\mathrm{k}\left[\mathbf{m}_\mathbf{f}\right] \tag{F.10}$$
$$=-2\mathbf{f}^\mathsf{T}\left((\mathbf{I}-\mathbb{E}^\mathrm{k}\left[\mathbf{A}_\mathrm{f}\right])^\mathsf{T}\mathbb{E}^\mathrm{k}\left[\mathbf{Q}_v^{-1}\right](\mathbf{I}-\mathbb{E}^\mathrm{k}\left[\mathbf{A}_\mathrm{m}\right]) + \mathrm{cov}^\mathrm{k}\left[\mathbf{A}_\mathrm{f}^\mathsf{T}\mathbf{L}_v^\mathsf{T}, \mathbf{A}_\mathrm{m}^\mathsf{T}\mathbf{L}_v^\mathsf{T}\right]\right)\mathbb{E}^\mathrm{k}\left[\mathbf{m}_\mathbf{f}\right], \tag{F.11}$$

where we use the notation for the cross-covariance $\mathrm{cov}^\mathrm{k}\left[\mathbf{X}^\mathsf{T},\mathbf{Y}^\mathsf{T}\right] = \mathbb{E}^\mathrm{k}\left[\mathbf{X}^\mathsf{T}\mathbf{Y} - \mathbb{E}^\mathrm{k}\left[\mathbf{X}\right]^\mathsf{T}\mathbb{E}^\mathrm{k}\left[\mathbf{Y}\right]\right]$.

From the above definitions, we deduce that $\mathrm{q}^\mathrm{k}(\mathbf{f}) \sim \mathcal{N}\left(\mathbf{f}\,\middle|\,\mathbb{E}^\mathrm{k}\left[\mathbf{f}\right], \mathrm{cov}^\mathrm{k}\left[\mathbf{f}\right]\right)$, since $\log \mathrm{q}^\mathrm{k}(\mathbf{f}) \propto -\frac{1}{2}\left(\mathbf{f}^\mathsf{T}\left(\mathrm{cov}^\mathrm{k}\left[\mathbf{f}\right]\right)^{-1}\mathbf{f} - 2\mathbf{f}^\mathsf{T}\left(\mathrm{cov}^\mathrm{k}\left[\mathbf{f}\right]\right)^{-1}\mathbb{E}^\mathrm{k}\left[\mathbf{f}\right] + \mathbb{E}^\mathrm{k}\left[\mathbf{f}\right]^\mathsf{T}\left(\mathrm{cov}^\mathrm{k}\left[\mathbf{f}\right]\right)^{-1}\mathbb{E}^\mathrm{k}\left[\mathbf{f}\right]\right)$, and by equating quadratic and linear terms in $\mathbf{f}$ we have that

$$\mathrm{cov}^\mathrm{k}\left[\mathbf{f}\right] = \left(\mathbb{E}^\mathrm{k}\left[\sigma_w^{-2}\right](\mathbb{E}^\mathrm{k}\left[\mathbf{H}\right]^\mathsf{T}\mathbb{E}^\mathrm{k}\left[\mathbf{H}\right] + \mathrm{cov}^\mathrm{k}\left[\mathbf{H}^\mathsf{T}\right]) + \|\mathbf{I}-\mathbb{E}^\mathrm{k}\left[\mathbf{A}_\mathrm{f}\right]\|^2_{\mathbf{Q}_v} + \mathrm{cov}^\mathrm{k}\left[\mathbf{A}_\mathrm{f}^\mathsf{T}\mathbf{L}_v^\mathsf{T}\right]\right)^{-1} \tag{F.12}$$

$$\mathbb{E}^\mathrm{k}\left[\mathbf{f}\right] = \mathrm{cov}^\mathrm{k}\left[\mathbf{f}\right]\left(\mathbb{E}^\mathrm{k}\left[\sigma_w^{-2}\right]\mathbb{E}^\mathrm{k}\left[\mathbf{H}\right]^\mathsf{T}\mathbf{g}\right.$$
$$\left. + \left((\mathbf{I}-\mathbb{E}^\mathrm{k}\left[\mathbf{A}_\mathrm{f}\right])^\mathsf{T}\mathbb{E}^\mathrm{k}\left[\mathbf{Q}_v^{-1}\right](\mathbf{I}-\mathbb{E}^\mathrm{k}\left[\mathbf{A}_\mathrm{m}\right]) + \mathrm{cov}^\mathrm{k}\left[\mathbf{A}_\mathrm{f}^\mathsf{T}\mathbf{L}_v^\mathsf{T}, \mathbf{A}_\mathrm{m}^\mathsf{T}\mathbf{L}_v^\mathsf{T}\right]\right)\mathbb{E}^\mathrm{k}\left[\mathbf{m}_\mathbf{f}\right]\right). \tag{F.13}$$

F.2 Distribution Approximation for $\mathbf{a}$

For the distribution $q^k(\mathbf{a}_f)$, using (5.25) (again ignoring terms independent of $\mathbf{a}_f$), we find:

$$\log q^k(\mathbf{a}_f) \propto \mathbb{E}\left[\log\left(p(\mathbf{f}\,|\,\mathbf{a}_f,\mu,\sigma_\nu)\,p(\mathbf{a}_f\,|\,\delta_a)\right)\right]_{q^k(\Theta_{\setminus a_f})} \tag{F.14}$$

$$\propto -\frac{1}{2}\left(\mathbb{E}\left[\|\boldsymbol{\nu}\|^2_{Q_\nu}\right]_{q^k(\mathbf{f},\mathbf{m}_f,\sigma_\nu)} + \mathbb{E}\left[\delta_a^{-1}\mathbf{a}_f^T\mathbf{a}_f\right]_{q^k(\delta_a)}\right), \tag{F.15}$$

To calculate the first expectation here, observe that $\boldsymbol{\nu}$ is independent in each block. Also, upon re-ordering $\mathbf{f}$ such that it is raster scanned one block at a time, $\mathbf{F}_a$ and $\mathbf{M}_a$ become block diagonal as $\mathbf{F}'_a$ and $\mathbf{M}'_a$. Thus we have

$$\mathbb{E}\left[\|\boldsymbol{\nu}\|^2_{Q_\nu}\right]_{q^k(\mathbf{f},\mathbf{m}_f,\sigma_\nu)} = \mathbb{E}\left[\|(\mathbf{f}^T-\mathbf{m}_f)-(\mathbf{F}_a-\mathbf{M}_a)\mathbf{a}_f\|^2_{Q_\nu}\right]_{q^k(\mathbf{f},\mathbf{m}_f,\sigma_\nu)} \tag{F.16}$$

$$= \sum_{r=1}^{R}\mathbb{E}\left[\sigma_{\nu_r}^{-2}\|(\mathbf{f}_r-\mathbf{m}_{f_r})-(\mathbf{F}_{a_r}-\mathbf{M}_{a_r})\mathbf{a}_{f_r}\|^2\right]_{q^k(\mathbf{f}_{r_+},\mu_r,\sigma_{\nu_r})} \tag{F.17}$$

$$= \sum_{r=1}^{R}\mathbb{E}^k\left[\sigma_{\nu_r}^2\right]^{-1}\Big(\mathbf{a}_{f_r}^T\mathbb{E}\left[\|\mathbf{F}_{a_r}-\mathbf{M}_{a_r}\|^2\right]_{q^k(\mathbf{f}_{r_+},\mu_r)}\mathbf{a}_{f_r}$$
$$-2\mathbf{a}_{f_r}^T\mathbb{E}\left[(\mathbf{F}_{a_r}-\mathbf{M}_{a_r})^T(\mathbf{f}_r-\mathbf{m}_{f_r})\right]_{q^k(\mathbf{f}_{r_+},\mu_r)} + \text{term independent of } \mathbf{a}_{f_r}\Big), \tag{F.18}$$

We see that the AR coefficients in each block are also independent. The expectations in (F.18) are then calculated as

$$\mathbb{E}\left[\|\mathbf{F}_{a_r}-\mathbf{M}_{a_r}\|^2\right]_{q^k(\mathbf{f}_r,\mu_r)} = \mathbb{E}\left[\mathbf{F}_{a_r}^T\mathbf{F}_{a_r}\right]_{q^k(\mathbf{f}_{r_+})} - 2\mathbb{E}\left[\mu_r\mathbf{1}_{L_a\times L_r}\mathbf{F}_{a_r}\right]_{q^k(\mathbf{f}_{r_+},\mu_r)}$$
$$+\mathbb{E}\left[\mu_r^2\right]_{q^k(\mu_r)}L_r\mathbf{1}_{L_a\times L_a} \tag{F.19}$$

$$= (\mathbb{E}^k[\mathbf{F}_{a_r}]^T\mathbb{E}^k[\mathbf{F}_{a_r}] + \mathrm{cov}^k\left[\mathbf{F}_{a_r}^T\right]) - 2\mathbb{E}^k[\mu_r]\,\mathbf{1}_{L_a\times L_r}\mathbb{E}^k[\mathbf{F}_{a_r}]$$
$$+(\mathbb{E}^k[\mu_r]^2+\mathrm{cov}^k[\mu_r])L_r\mathbf{1}_{L_a\times L_a} \tag{F.20}$$

$$= \|\mathbb{E}^k[\mathbf{F}_{a_r}]-\mathbb{E}^k[\mathbf{M}_{a_r}]\|^2 + \mathrm{cov}^k[\mathbf{F}_{a_r}] + \mathrm{cov}^k[\mu_r]\,L_r\mathbf{1}_{L_a\times L_a} \tag{F.21}$$

and

$$\mathbb{E}\left[(\mathbf{F}_{a_r}-\boldsymbol{M}_{a_r})^{\mathsf{T}}(\mathbf{f}_r-\boldsymbol{m}_{f_r})\right]_{q^k(\mathbf{f}_{r_+},\mu_r)} = \mathbb{E}\left[\mathbf{F}_{a_r}^{\mathsf{T}}\mathbf{f}_r\right]_{q^k(\mathbf{f}_{r_+})} - \mathbb{E}\left[\mu_r\mathbf{F}_{a_r}^{\mathsf{T}}\mathbf{1}_{L_r}\right]_{q^k(\mathbf{f}_{r_+},\mu_r)}$$
$$- \mathbb{E}\left[\mu_r\mathbf{1}_{L_a\times L_r}\mathbf{f}_r\right]_{q^k(\mathbf{f}_{r_+},\mu_r)} + \mathbb{E}\left[\mu_r^2\right]_{q^k(\mu_r)} L_r\mathbf{1}_{L_a} \tag{F.22}$$

$$= (\mathbb{E}^k[\mathbf{F}_{a_r}]^{\mathsf{T}}\mathbb{E}^k[\mathbf{f}_r] + \mathrm{cov}^k\left[\mathbf{F}_{a_r}^{\mathsf{T}},\mathbf{f}_r^{\mathsf{T}}\right]) - \mathbb{E}^k[\mu_r]\,\mathbb{E}^k[\mathbf{F}_{a_r}]^{\mathsf{T}}\mathbf{1}_{L_r}$$
$$- \mathbb{E}^k[\mu_r]\,\mathbf{1}_{L_a\times L_r}\mathbb{E}^k[\mathbf{f}_r] + (\mathbb{E}^k[\mu_r]^2 + \mathrm{cov}^k[\mu_r])L_r\mathbf{1}_{L_a} \tag{F.23}$$

$$= (\mathbb{E}^k[\mathbf{F}_{a_r}] - \mathbb{E}^k[\boldsymbol{M}_{a_r}])^{\mathsf{T}}(\mathbb{E}^k[\mathbf{f}_r] - \mathbb{E}^k[\boldsymbol{m}_{f_r}])$$
$$+ \mathrm{cov}^k\left[\mathbf{F}_{a_r}^{\mathsf{T}},\mathbf{f}_r^{\mathsf{T}}\right] + \mathrm{cov}^k[\mu_r]\,L_r\mathbf{1}_{L_a}. \tag{F.24}$$

Therefore we deduce that $q^k(\boldsymbol{a}_{f_r}) \sim \mathcal{N}\left(\boldsymbol{a}_{f_r} \,\middle|\, \mathbb{E}^k[\boldsymbol{a}_{f_r}], \mathrm{cov}^k[\boldsymbol{a}_{f_r}]\right)$, with

$$\mathrm{cov}^k[\boldsymbol{a}_{f_r}] = \left(\|\mathbb{E}^k[\mathbf{F}_{a_r}] - \mathbb{E}^k[\boldsymbol{M}_{a_r}]\|^2 + \mathrm{cov}^k[\mathbf{F}_{a_r}] + \mathrm{cov}^k[\mu_r]\,L_r\mathbf{1}_{L_a\times L_a} + \delta_a^{-1}\mathbf{I}_{L_a}\right)^{-1} \tag{F.25}$$

$$\mathbb{E}^k[\boldsymbol{a}_{f_r}] = \mathrm{cov}^k[\boldsymbol{a}_{f_r}]\left((\mathbb{E}^k[\mathbf{F}_{a_r}] - \mathbb{E}^k[\boldsymbol{M}_{a_r}])^{\mathsf{T}}(\mathbb{E}^k[\mathbf{f}_r] - \mathbb{E}^k[\boldsymbol{m}_{f_r}]) + \mathrm{cov}^k\left[\mathbf{F}_{a_r}^{\mathsf{T}},\mathbf{f}_r^{\mathsf{T}}\right] + \mathrm{cov}^k[\mu_r]\,L_r\mathbf{1}_{L_a}\right). \tag{F.26}$$

F.3 Distribution Approximation for $\mathfrak{m}$

For the distribution $q^k(\boldsymbol{m}_{\mathbf{f}})$, using (5.25) (again ignoring terms independent of $\boldsymbol{m}_{\mathbf{f}}$), we find:

$$\log q^k(\boldsymbol{m}_{\mathbf{f}}) \propto \mathbb{E}\left[\log\left(p(\mathbf{f}\,|\,\boldsymbol{m}_{\mathbf{f}},\boldsymbol{a}_f,\sigma_v)\right)\right]_{q^k(\boldsymbol{\Theta}_{\setminus\boldsymbol{m}_{\mathbf{f}}})} \tag{F.27}$$

$$\propto -\frac{1}{2}\left(\mathbb{E}\left[\|\boldsymbol{v}\|^2_{\mathbf{Q}_v}\right]_{q^k(\mathbf{f},\boldsymbol{a}_f,\sigma_v)}\right), \tag{F.28}$$

In a similar manner as for the derivation for $q^k(\boldsymbol{a}_f)$, we may here split $\boldsymbol{v}$ into its constituent blocks, giving

$$\mathbb{E}\left[\|\boldsymbol{v}\|^2_{\mathbf{Q}_v}\right]_{q^k(\mathbf{f},\boldsymbol{a}_f,\sigma_v)} = \sum_{r=1}^{R}\mathbb{E}\left[\sigma_{v_r}^{-2}\|(\mathbf{f}_r-\boldsymbol{m}_{f_r}) - (\mathbf{F}_{a_r}-\boldsymbol{M}_{a_r})\boldsymbol{a}_{f_r}\|^2\right]_{q^k(\mathbf{f},\boldsymbol{a}_f,\sigma_v)} \tag{F.29}$$

$$= \sum_{r=1}^{R} \mathbb{E}^k\left[\sigma_{\nu_r}^2\right]^{-1}\left(\mathbb{E}\left[\|(\mathbf{I}_+ - \mathbf{A}_{\mathbf{f}_{r_+}})\mathbf{f}_{r_+} - (\mathbf{I}-\mathbf{A}_{m_r})\mathbf{m}_{\mathbf{f}_r}\|^2\right]_{q^k(\mathbf{f}_{r_+},\mathbf{a}_{\mathbf{f}_r})}\right) \tag{F.30}$$

$$\begin{aligned}= \sum_{r=1}^{R} \mathbb{E}^k\left[\sigma_{\nu_r}^2\right]^{-1}\Bigg(&\mathbb{E}\left[\mathbf{m}_{\mathbf{f}_r}^T(\mathbf{I}-\mathbf{A}_{m_r})^T(\mathbf{I}-\mathbf{A}_{m_r})\mathbf{m}_{\mathbf{f}_r}\right]_{q^k(\mathbf{a}_{\mathbf{f}_r})} \\ &- 2\mathbb{E}\left[\mathbf{m}_{\mathbf{f}_r}^T(\mathbf{I}-\mathbf{A}_{m_r})^T(\mathbf{I}_+ - \mathbf{A}_{\mathbf{f}_{r_+}})\mathbf{f}_{r_+}\right]_{q^k(\mathbf{f}_{r_+},\mathbf{a}_{\mathbf{f}_r})} + \text{term independent of } \mathbf{m}_{\mathbf{f}_r}\Bigg).\end{aligned} \tag{F.31}$$

The first term of (F.31) is found as

$$\mathbb{E}\left[\|(\mathbf{I}-\mathbf{A}_{m_r})\mathbf{1}_{L_r}\mu_r\|^2\right]_{q^k(\mathbf{a}_{\mathbf{f}_r})} = \mu_r^2\mathbf{1}_{L_r}^T\left(\mathbf{I} - 2\mathbb{E}^k\left[\mathbf{A}_{m_r}\right]^T + \mathbb{E}^k\left[\mathbf{A}_{m_r}^T\mathbf{A}_{m_r}\right]\right)\mathbf{1}_{L_r} \tag{F.32}$$

$$\mu_r^2\mathbf{1}_{L_r}^T\left(\|\mathbf{I} - \mathbb{E}^k\left[\mathbf{A}_{m_r}\right]\|^2 + \operatorname{cov}^k\left[\mathbf{A}_{m_r}^T\right]\right)\mathbf{1}_{L_r} \tag{F.33}$$

However observe that since μ_r is scalar we can simplify this expression. $\mathbf{A}_{m_r}$ is BCCB; in the 1D causal case it is circulant with first column given by $[1, -a_{r_1}, \cdots, -a_{r_p}, 0, \cdots, 0]^T$, which gives $(\mathbf{I}-\mathbf{A}_{m_r})\mathbf{1}_{L_r} = (1 - \sum_{k=1}^{P} a_{r_k})\mathbf{1}_{L_r}$, and this expression also holds for the 2D BCCB case with the sum over the L_a valid AR coefficients. Thus:

$$\mathbb{E}\left[\|(\mathbf{I}-\mathbf{A}_{m_r})\mathbf{1}_{L_r}\mu_r\|^2\right]_{q^k(\mathbf{a}_{\mathbf{f}_r})} = \mu_r^2 L_r\mathbb{E}\left[(1-\sum_{k=1}^{L_a} a_{r_k})^2\right]_{q^k(\mathbf{a}_{\mathbf{f}_r})} \tag{F.34}$$

$$= \mu_r^2 L_r\left(1 - 2\sum_{k=1}^{L_a}\mathbb{E}^k\left[a_{r_k}\right] + \mathbb{E}^k\left[\left(\sum_{k=1}^{L_a} a_{r_k}\right)^2\right]\right) \tag{F.35}$$

$$= \mu_r^2 L_r(1 - 2\cdot\mathbf{1}_{L_a}^T\mathbb{E}^k\left[\mathbf{a}_{\mathbf{f}_r}\right] + \mathbf{1}_{L_a}^T\mathbb{E}^k\left[\mathbf{a}_{\mathbf{f}_r}\mathbf{a}_{\mathbf{f}_r}^T\right]\mathbf{1}_{L_a}) \tag{F.36}$$

$$= \mu_r^2 L_r(1 - 2\cdot\mathbf{1}_{L_a}^T\mathbb{E}^k\left[\mathbf{a}_{\mathbf{f}_r}\right] + \mathbf{1}_{L_a}^T\left(\mathbb{E}^k\left[\mathbf{a}_{\mathbf{f}_r}\right]\mathbb{E}^k\left[\mathbf{a}_{\mathbf{f}_r}\right]^T + \operatorname{cov}^k\left[\mathbf{a}_{\mathbf{f}_r}\right]\right)\mathbf{1}_{L_a}) \tag{F.37}$$

$$= \mu_r^2 L_r\left(\|1 - \mathbf{1}_{L_a}^T\mathbb{E}^k\left[\mathbf{a}_{\mathbf{f}_r}\right]\|^2 + \mathbf{1}_{L_a}^T\operatorname{cov}^k\left[\mathbf{a}_{\mathbf{f}_r}\right]\mathbf{1}_{L_a}\right) \tag{F.38}$$

The second term of (F.31) is similarly found as

$$\mathbb{E}\left[\mathbf{m}_{\mathbf{f}_r}^T(\mathbf{I}-\mathbf{A}_{m_r})^T(\mathbf{I}_+ - \mathbf{A}_{\mathbf{f}_{r_+}})\mathbf{f}_{r_+}\right]_{q^k(\mathbf{f}_{r_+},\mathbf{a}_{\mathbf{f}_r})} \tag{F.39}$$

$$= \mu_r\mathbf{1}_{L_r}^T\left(\mathbf{I}_+ - \mathbb{E}^k\left[\mathbf{A}_{m_r}\right]^T - \mathbb{E}^k\left[\mathbf{A}_{\mathbf{f}_{r_+}}\right] + \mathbb{E}^k\left[\mathbf{A}_{m_r}^T\mathbf{A}_{\mathbf{f}_{r_+}}\right]\right)\mathbb{E}^k\left[\mathbf{f}_{r_+}\right] \tag{F.40}$$

$$= \mu_r \mathbf{1}_{L_r}^\top \left((\mathbf{I} - \mathbb{E}^k [\mathbf{A}_{m_r}])^\top (\mathbf{I}_+ - \mathbb{E}^k \left[\mathbf{A}_{f_{r_+}} \right]) + \mathrm{cov}^k \left[\mathbf{A}_{m_r}^\top, \mathbf{A}_{f_{r_+}}^\top \right] \right) \mathbb{E}^k \left[\mathbf{f}_{r_+} \right] \tag{F.41}$$

$$= \mu_r \left((1 - \mathbf{1}_{L_a}^\top \mathbb{E}^k [\mathbf{a}_{f_r}]) \mathbf{1}_{L_r}^\top (\mathbf{I}_+ - \mathbb{E}^k \left[\mathbf{A}_{f_{r_+}} \right]) + \mathrm{cov}^k \left[\mathbf{A}_{m_r}^\top, \mathbf{A}_{f_{r_+}}^\top \right] \right) \mathbb{E}^k \left[\mathbf{f}_{r_+} \right] \tag{F.42}$$

or

$$\mu_r \mathbb{E} \left[\mathbf{1}_{L_r}^\top (\mathbf{I} - \mathbf{A}_{m_r})^\top (\mathbf{I}_+ - \mathbf{A}_{f_{r_+}}) \mathbf{f}_{r_+} \right]_{q^k(\mathbf{f}_{r_+}, \mathbf{a}_{f_r})} = \mu_r \mathbb{E} \left[(1 - \sum_{k=1}^{L_a} a_{r_k}) \mathbf{1}_{L_r}^\top (\mathbf{I}_+ - \mathbf{A}_{f_{r_+}}) \right]_{q^k(\mathbf{a}_{f_r})} \mathbb{E}^k \left[\mathbf{f}_{r_+} \right] \tag{F.43}$$

Note however that the term $\mathbf{1}_{L_r}^\top (\mathbf{I}_+ - \mathbf{A}_{f_{r_+}})$ does not simplify in the same manner, due to the boundary conditions used in $\mathbf{A}_{f_{r_+}}$.

We then deduce that $q^k(\mu_r) \sim \mathcal{N}\left(\mu_r \,\middle|\, \mathbb{E}^k[\mu_r], \mathrm{cov}^k[\mu_r]\right)$, with

$$\mathrm{cov}^k[\mu_r] = \left(L_r \left(\| 1 - \mathbf{1}_{L_a}^\top \mathbb{E}^k [\mathbf{a}_{f_r}] \|^2 + \mathbf{1}_{L_a}^\top \mathrm{cov}^k [\mathbf{a}_{f_r}] \mathbf{1}_{L_a} \right) \right)^{-1} \tag{F.44}$$

$$\mathbb{E}^k[\mu_r] = \mathrm{cov}^k[\mu_r] \left(\mathbb{E} \left[(1 - \sum_{k=1}^{L_a} a_{r_k}) \mathbf{1}_{L_r}^\top (\mathbf{I} - \mathbf{A}_{f_{r_+}}) \right]_{q^k(\mathbf{a}_{f_r})} \mathbb{E}^k \left[\mathbf{f}_{r_+} \right] \right) \tag{F.45}$$

$$= \frac{\left((1 - \mathbf{1}_{L_a}^\top \mathbb{E}^k [\mathbf{a}_{f_r}]) \mathbf{1}_{L_r}^\top (\mathbf{I}_+ - \mathbb{E}^k \left[\mathbf{A}_{f_{r_+}} \right]) + \mathbf{1}_{L_r}^\top \mathrm{cov}^k \left[\mathbf{A}_{m_r}^\top, \mathbf{A}_{f_{r_+}}^\top \right] \right) \mathbb{E}^k \left[\mathbf{f}_{r_+} \right]}{L_r \left(\| 1 - \mathbf{1}_{L_a}^\top \mathbb{E}^k [\mathbf{a}_{f_r}] \|^2 + \mathbf{1}_{L_a}^\top \mathrm{cov}^k [\mathbf{a}_{f_r}] \mathbf{1}_{L_a} \right)} \tag{F.46}$$

F.4 Distribution Approximation for $\mathbf{h}$

For the distribution $q^k(\mathbf{h})$, using (5.25) (again ignoring terms independent of $\mathbf{h}$), we find:

$$\log q^k(\mathbf{h}) \propto \mathbb{E} \left[\log \left(p(\mathbf{g} \mid \boldsymbol{\Theta}) p(\mathbf{h} \mid \sigma_u) \right) \right]_{q^k(\boldsymbol{\Theta}_{\setminus \mathbf{h}})} \tag{F.47}$$

$$\propto -\frac{1}{2} \left(\mathbb{E} \left[\sigma_w^{-2} \| \mathbf{g} - \mathbf{F}\mathbf{h} \|^2 \right]_{q^k(\mathbf{f}, \sigma_w)} + \mathbb{E} \left[\| \mathbf{C}\mathbf{h} \|_{\mathbf{Q}_u}^2 \right]_{q^k(\mathbf{u})} \right) \tag{F.48}$$

$$\propto -\frac{1}{2} \left(\mathbb{E}^k \left[\sigma_w^{-2} \right] \| \mathbf{g} - \mathbb{E}^k [\mathbf{F}] \mathbf{h} \|^2 + \mathbf{h}^\top \mathrm{cov}^k \left[\mathbf{F}^\top \right] \mathbf{h} + \| \mathbf{C}\mathbf{h} \|_{\mathbb{E}^k[\mathbf{Q}_u]}^2 \right), \tag{F.49}$$

using similar steps as for those in the derivation of $q^k(\mathbf{f})$. Therefore we have $q^k(\mathbf{h}) \sim \mathcal{N}\left(\mathbf{h} \middle| \mathbb{E}^k[\mathbf{h}], \text{cov}^k[\mathbf{h}]\right)$, with

$$\text{cov}^k[\mathbf{h}] = \left(\mathbb{E}^k\left[\sigma_w^{-2}\right](\mathbb{E}^k[\mathbf{F}]^T\mathbb{E}^k[\mathbf{F}] + \text{cov}^k\left[\mathbf{F}^T\right]) + \mathbf{C}^T\mathbb{E}^k[\mathbf{Q}_u]^{-1}\mathbf{C}\right)^{-1} \tag{F.50}$$

$$\mathbb{E}^k[\mathbf{h}] = \text{cov}^k[\mathbf{h}]\left(\mathbb{E}^k\left[\sigma_w^{-2}\right]\mathbb{E}^k[\mathbf{F}]^T\mathbf{g}\right) \tag{F.51}$$

F.5 Distribution Approximation for σ_w

For the distribution $q^k(\sigma_w)$, using (5.25) (again ignoring terms independent of σ_w), we find:

$$\log q^k(\sigma_w) \propto \mathbb{E}\left[\log\left(p(\mathbf{g}|\boldsymbol{\Theta})p(\sigma_w)\right)\right]_{q(\boldsymbol{\Theta}_{\setminus\sigma_w})} \tag{F.52}$$

$$\propto \log(p(\sigma_w)) - \frac{1}{2}L\log(\sigma_w^2) + \mathbb{E}\left[\sigma_w^{-2}\boldsymbol{w}^T\boldsymbol{w}\right]_{q^k(\mathbf{f},\mathbf{h})} \tag{F.53}$$

$$\Rightarrow q^k(\sigma_w) = \prod_{r=1}^{R_h}\mathcal{IG}\left(\sigma_{u_r}^2 \middle| \alpha_w + \frac{L_r}{2}, \beta_w + \frac{1}{2}\mathbb{E}\left[\boldsymbol{w}^T\boldsymbol{w}\right]_{q(\mathbf{f},\mathbf{h})}\right) \tag{F.54}$$

This is in the form of a product of inverse Gamma distributions over the variances in each region of the PSF. Hence to find $q^{k+1}\left(\sigma_{u_r}^2\right)$, we now just require the expectation

$$\mathbb{E}\left[\boldsymbol{w}^T\boldsymbol{w}\right]_{q(\mathbf{f},\mathbf{h})} = \mathbb{E}\left[\sigma_w^{-2}\|\mathbf{g}-\mathbf{F}\mathbf{h}\|^2\right]_{q^k(\mathbf{f},\mathbf{h})} \tag{F.55}$$

$$= \mathbb{E}^k\left[\sigma_w^2\right]^{-1}\left(\mathbf{g}^T\mathbf{g} - 2\mathbf{g}^T\mathbb{E}^k[\mathbf{F}]\mathbb{E}^k[\mathbf{h}] + \mathbb{E}\left[\mathbf{h}^T\mathbf{F}^T\mathbf{F}\mathbf{h}\right]_{q^k(\mathbf{f},\mathbf{h})}\right). \tag{F.56}$$

To find this we first consider expanding the product

$$\text{tr}\left[\text{cov}^k\left[\mathbf{F}^T\right]\text{cov}^k[\mathbf{h}]\right] = \text{tr}\left[\left(\mathbb{E}^k\left[\mathbf{F}^T\mathbf{F}\right] - \mathbb{E}^k[\mathbf{F}]^T\mathbb{E}^k[\mathbf{F}]\right)\left(\mathbb{E}^k\left[\mathbf{h}\mathbf{h}^T\right] - \mathbb{E}^k[\mathbf{h}]\mathbb{E}^k[\mathbf{h}]^T\right)\right] \tag{F.57}$$

$$= \mathbb{E}\left[\text{tr}\left[\mathbf{F}^T\mathbf{F}\mathbf{h}\mathbf{h}^T\right]\right]_{q^k(\mathbf{f},\mathbf{h})} - \text{tr}\left[\mathbb{E}^k\left[\mathbf{F}^T\mathbf{F}\right]\mathbb{E}^k[\mathbf{h}]\mathbb{E}^k[\mathbf{h}]^T\right]$$
$$- \text{tr}\left[\mathbb{E}^k[\mathbf{F}]^T\mathbb{E}^k[\mathbf{F}]\mathbb{E}^k\left[\mathbf{h}\mathbf{h}^T\right]\right] + \text{tr}\left[\mathbb{E}^k[\mathbf{F}]^T\mathbb{E}^k[\mathbf{F}]\mathbb{E}^k[\mathbf{h}]\mathbb{E}^k[\mathbf{h}]^T\right] \tag{F.58}$$

$$= \mathbb{E}\left[\mathbf{h}^T\mathbf{F}^T\mathbf{F}\mathbf{h}\right]_{q^k(\mathbf{f},\mathbf{h})} - \text{tr}\left[\mathbb{E}\left[\mathbb{E}^k[\mathbf{h}]^T\mathbf{F}^T\mathbf{F}\mathbb{E}^k[\mathbf{h}]\right]_{q^k(\mathbf{f})}\right]$$
$$- \text{tr}\left[\mathbb{E}^k[\mathbf{F}]^T\mathbb{E}^k[\mathbf{F}]\left(\text{cov}^k[\mathbf{h}] + \mathbb{E}^k[\mathbf{h}]\mathbb{E}^k\left[\mathbf{h}^T\right]\right)\right] + \mathbb{E}^k[\mathbf{h}]^T\mathbb{E}^k[\mathbf{F}]^T\mathbb{E}^k[\mathbf{F}]\mathbb{E}^k[\mathbf{h}] \tag{F.59}$$

and the second term in this expression is further expanded as

$$\begin{aligned}
\mathrm{tr}\left[\mathbb{E}\left[\mathbb{E}^{k}\left[\mathbf{h}\right]^{T}\mathbf{F}^{T}\mathbf{F}\mathbb{E}^{k}\left[\mathbf{h}\right]\right]_{q^{k}(\mathbf{f})}\right] &= \mathrm{tr}\left[\mathbb{E}\left[\mathbf{f}^{T}\mathbb{E}^{k}\left[\mathbf{H}\right]^{T}\mathbb{E}^{k}\left[\mathbf{H}\right]\mathbf{f}\right]_{q^{k}(\mathbf{f})}\right] && \text{(F.60)}\\
&= \mathrm{tr}\left[\mathbb{E}^{k}\left[\mathbf{H}\right]^{T}\mathbb{E}^{k}\left[\mathbf{H}\right]\mathbb{E}^{k}\left[\mathbf{f}\mathbf{f}^{T}\right]\right] && \text{(F.61)}\\
&= \mathrm{tr}\left[\mathbb{E}^{k}\left[\mathbf{H}\right]^{T}\mathbb{E}^{k}\left[\mathbf{H}\right]\left(\mathbb{E}^{k}\left[\mathbf{f}\right]\mathbb{E}^{k}\left[\mathbf{f}^{T}\right]+\mathrm{cov}^{k}\left[\mathbf{f}\right]\right)\right] && \text{(F.62)}\\
&= \mathbb{E}^{k}\left[\mathbf{f}^{T}\right]\mathbb{E}^{k}\left[\mathbf{H}\right]^{T}\mathbb{E}^{k}\left[\mathbf{H}\right]\mathbb{E}^{k}\left[\mathbf{f}\right]+\mathrm{tr}\left[\mathbb{E}^{k}\left[\mathbf{H}\right]^{T}\mathbb{E}^{k}\left[\mathbf{H}\right]\mathrm{cov}^{k}\left[\mathbf{f}\right]\right] && \text{(F.63)}
\end{aligned}$$

and therefore substituting this into (F.59) and rearranging gives

$$\begin{aligned}
\mathbb{E}\left[\mathbf{h}^{T}\mathbf{F}^{T}\mathbf{F}\mathbf{h}\right]_{q^{k}(\mathbf{f},\mathbf{h})} &= \mathrm{tr}\left[\mathrm{cov}^{k}\left[\mathbf{F}^{T}\right]\mathrm{cov}^{k}\left[\mathbf{h}\right]\right]+\mathbb{E}^{k}\left[\mathbf{f}^{T}\right]\mathbb{E}^{k}\left[\mathbf{H}\right]^{T}\mathbb{E}^{k}\left[\mathbf{H}\right]\mathbb{E}^{k}\left[\mathbf{f}\right]\\
&\quad+\mathrm{tr}\left[\mathbb{E}^{k}\left[\mathbf{H}\right]^{T}\mathbb{E}^{k}\left[\mathbf{H}\right]\mathrm{cov}^{k}\left[\mathbf{f}\right]\right]+\mathrm{tr}\left[\mathbb{E}^{k}\left[\mathbf{F}\right]^{T}\mathbb{E}^{k}\left[\mathbf{F}\right]\mathrm{cov}^{k}\left[\mathbf{h}\right]\right] && \text{(F.64)}\\
&\quad+\mathbb{E}^{k}\left[\mathbf{h}\right]^{T}\mathbb{E}^{k}\left[\mathbf{F}\right]^{T}\mathbb{E}^{k}\left[\mathbf{F}\right]\mathbb{E}^{k}\left[\mathbf{h}\right]-\mathbb{E}^{k}\left[\mathbf{h}\right]^{T}\mathbb{E}^{k}\left[\mathbf{F}\right]^{T}\mathbb{E}^{k}\left[\mathbf{F}\right]\mathbb{E}^{k}\left[\mathbf{h}\right]
\end{aligned}$$

and the last two terms cancel. Hence

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{w}^{T}\boldsymbol{w}\right]_{q(\mathbf{f},\mathbf{h})} &= \mathbb{E}^{k}\left[\sigma_{w}^{2}\right]^{-1}\left(\mathbf{g}^{T}\mathbf{g}-2\mathbf{g}^{T}\mathbb{E}^{k}\left[\mathbf{F}\right]\mathbb{E}^{k}\left[\mathbf{h}\right]+\mathbb{E}\left[\mathbf{h}^{T}\mathbf{F}^{T}\mathbf{F}\mathbf{h}\right]_{q^{k}(\mathbf{f},\mathbf{h})}\right) && \text{(F.65)}\\
&= \mathbb{E}^{k}\left[\sigma_{w}^{-2}\right]\|\mathbf{g}-\mathbb{E}^{k}\left[\mathbf{H}\right]\mathbb{E}^{k}\left[\mathbf{f}\right]\|^{2}+\mathrm{tr}\left[\mathrm{cov}^{k}\left[\mathbf{F}^{T}\right]\mathrm{cov}^{k}\left[\mathbf{h}\right]\right]\\
&\quad+\mathrm{tr}\left[\mathbb{E}^{k}\left[\mathbf{H}\right]^{T}\mathbb{E}^{k}\left[\mathbf{H}\right]\mathrm{cov}^{k}\left[\mathbf{f}\right]\right]+\mathrm{tr}\left[\mathbb{E}^{k}\left[\mathbf{F}\right]^{T}\mathbb{E}^{k}\left[\mathbf{F}\right]\mathrm{cov}^{k}\left[\mathbf{h}\right]\right] && \text{(F.66)}
\end{aligned}$$

F.6 Distribution Approximation for δ_{a_f}

In a similar manner to that for σ_w, we have for δ_a:

$$\begin{aligned}
\log q^{k}\left(\delta_{a}\right) &\propto \mathbb{E}\left[\log\left(p\left(\boldsymbol{a}_{f}\,|\,\delta_{a}\right)p\left(\delta_{a}\right)\right)\right]_{q(\boldsymbol{\Theta}_{\setminus\delta_{a}})} && \text{(F.67)}\\
&\propto \log(p\left(\delta_{a}\right))-\frac{1}{2}P\log(\delta_{a})+\mathbb{E}\left[\delta_{a}^{-1}\boldsymbol{a}_{f}^{T}\boldsymbol{a}_{f}\right]_{q(\boldsymbol{a}_{f})} && \text{(F.68)}\\
\Rightarrow q^{k}\left(\delta_{a}\right) &= \mathcal{IG}\left(\delta_{a}\,\middle|\,\alpha_{\delta}+\frac{P}{2},\beta_{\delta}+\frac{1}{2}\mathbb{E}\left[\boldsymbol{a}_{f}^{T}\boldsymbol{a}_{f}\right]_{q(\boldsymbol{a}_{f})}\right) && \text{(F.69)}\\
\Rightarrow q^{k}\left(\delta_{a}\right) &= \mathcal{IG}\left(\delta_{a}\,\middle|\,\alpha_{\delta}+\frac{P}{2},\beta_{\delta}+\frac{1}{2}\left(\mathbb{E}^{k}\left[\boldsymbol{a}_{f}\right]^{T}\mathbb{E}^{k}\left[\boldsymbol{a}_{f}\right]+\mathrm{tr}\left[\mathrm{cov}^{k}\left[\boldsymbol{a}_{f}\right]\right]\right)\right) && \text{(F.70)}
\end{aligned}$$

F.7 Distribution Approximation for σ_ν

For the distribution $q^k(\boldsymbol{\sigma}_\nu)$, using (5.25) (again ignoring terms independent of $\boldsymbol{\sigma}_\nu$), we find:

$$\log q^k(\boldsymbol{\sigma}_\nu) \propto \mathbb{E}\left[\log\left(p(\mathbf{f}\,|\,\boldsymbol{\sigma}_\nu)\,p(\boldsymbol{\sigma}_\nu)\right)\right]_{q(\boldsymbol{\Theta}_{\setminus\sigma_\nu})} \tag{F.71}$$

$$\propto \log(p(\boldsymbol{\sigma}_\nu)) - \frac{1}{2}\sum_{r=1}^{R} L_r \log(\sigma_{\nu_r}^2) + \mathbb{E}\left[\sigma_{\nu_r}^{-2}\boldsymbol{\nu}_r^T\boldsymbol{\nu}_r\right]_{q(\mathbf{f},\mathbf{m}_\mathbf{f},\mathbf{a}_\mathbf{f})} \tag{F.72}$$

$$\Rightarrow q^k(\boldsymbol{\sigma}_\nu) = \prod_{r=1}^{R} \mathcal{IG}\left(\sigma_{\nu_r}^2 \,\middle|\, \alpha_{\nu_r} + \frac{L_r}{2}, \beta_{\nu_r} + \frac{1}{2}\mathbb{E}\left[\boldsymbol{\nu}_r^T\boldsymbol{\nu}_r\right]_{q(\mathbf{f}_{r_+},\mu_r,\mathbf{a}_{f_r})}\right) \tag{F.73}$$

This is in the form of a product of inverse Gamma distributions over the variances in each region (we can again split into regions due to the independence of $\sigma_{\nu_r}^2$ in each region). Hence to find $q^{k+1}(\sigma_{\nu_r}^2)$, we now just require the expectation

$$\mathbb{E}\left[\boldsymbol{\nu}_r^T\boldsymbol{\nu}_r\right]_{q(\mathbf{f},\mathbf{m}_\mathbf{f},\mathbf{a}_\mathbf{f})} = \mathbb{E}\left[\|(\mathbf{I}_+ - \mathbf{A}_{f_{r_+}})\mathbf{f}_{r_+} - (\mathbf{I} - \mathbf{A}_{m_r})\mathbf{m}_{f_r}\|^2\right]_{q(\mathbf{f}_{r_+},\mu_r,\mathbf{a}_{f_r})} \tag{F.74}$$

$$= \mathbb{E}\left[\mathbf{m}_{f_r}^T(\mathbf{I} - \mathbf{A}_{m_r})^T(\mathbf{I} - \mathbf{A}_{m_r})\mathbf{m}_{f_r}\right]_{q^k(\mathbf{a}_{f_r},\mu_r)}$$

$$- 2\mathbb{E}\left[\mathbf{m}_{f_r}^T(\mathbf{I} - \mathbf{A}_{m_r})^T(\mathbf{I}_+ - \mathbf{A}_{f_{r_+}})\mathbf{f}_{r_+}\right]_{q^k(\mathbf{f}_{r_+},\mu_r,\mathbf{a}_{f_r})} \tag{F.75}$$

$$+ \mathbb{E}\left[\mathbf{f}_{r_+}^T(\mathbf{I} - \mathbf{A}_{f_{r_+}})^T(\mathbf{I}_+ - \mathbf{A}_{f_{r_+}})\mathbf{f}_{r_+}\right]_{q^k(\mathbf{f}_{r_+},\mathbf{a}_{f_r})}.$$

$$= (\mathbb{E}^k[\mu_r]^2 + \mathrm{cov}^k[\mu_r])\mathbf{1}_{L_r}^T\left(\|\mathbf{I} - \mathbb{E}^k[\mathbf{A}_{m_r}]\|^2 + \mathrm{cov}^k\left[\mathbf{A}_{m_r}^T\right]\right)\mathbf{1}_{L_r}$$

$$- 2\mathbb{E}^k[\mu_r]\,\mathbf{1}_{L_r}^T\left((\mathbf{I} - \mathbb{E}^k[\mathbf{A}_{m_r}])^T(\mathbf{I}_+ - \mathbb{E}^k\left[\mathbf{A}_{f_{r_+}}\right]) + \mathrm{cov}^k\left[\mathbf{A}_{m_r}^T, \mathbf{A}_{f_{r_+}}^T\right]\right)\mathbb{E}^k\left[\mathbf{f}_{r_+}\right]$$

$$+ \mathbb{E}\left[\mathbf{f}_{r_+}^T\mathbf{f}_{r_+} - 2\mathbf{f}_{r_+}^T\mathbf{A}_{f_{r_+}}\mathbf{f}_{r_+} + \mathbf{f}_{r_+}^T\mathbf{A}_{f_{r_+}}^T\mathbf{A}_{f_{r_+}}\mathbf{f}_{r_+}\right]_{q^k(\mathbf{f}_{r_+},\mathbf{a}_{f_r})}. \tag{F.76}$$

The first two expectations in this expression are found simply using the results for the distribution of $q^k(\mathbf{m}_{f_r})$. Becuase the third term includes an expectation of a quadratic in $\mathbf{A}_{f_{r_+}}\mathbf{f}_{r_+}$ we cannot reuse previous results here; this term is found as:

$$\mathbb{E}^k\left[\mathbf{f}_{r_+}^T\mathbf{f}_{r_+}\right] - 2\mathbb{E}\left[\mathbf{a}_{f_r}^T\mathbf{F}_{a_r}^T\mathbf{f}_{r_+}\right]_{q^k(\mathbf{f}_{r_+},\mathbf{a}_{f_r})} + \mathbb{E}\left[\mathbf{f}_{r_+}^T\mathbf{A}_{f_{r_+}}^T\mathbf{A}_{f_{r_+}}\mathbf{f}_{r_+}\right]_{q^k(\mathbf{f}_{r_+},\mathbf{a}_{f_r})} \tag{F.77}$$

$$
\begin{aligned}
&= \mathbb{E}^{k}\left[\mathbf{f}_{r_+}\right]^{T}\mathbb{E}^{k}\left[\mathbf{f}_{r_+}\right] + \mathrm{tr}\left[\mathrm{cov}^{k}\left[\mathbf{f}_{r_+}\right]\right] - 2\mathbb{E}^{k}\left[\boldsymbol{\alpha}_{f_r}\right]^{T}\left(\mathbb{E}^{k}\left[\mathbf{F}_{a_r}^{T}\right]\mathbb{E}^{k}\left[\mathbf{f}_{r_+}\right] + \mathrm{cov}^{k}\left[\mathbf{F}_{a_r}^{T}, \mathbf{f}_{r_+}^{T}\right]\right) \\
&+ \Bigg(\mathrm{tr}\left[\mathrm{cov}^{k}\left[\mathbf{F}_{a_r}^{T}\right]\mathrm{cov}^{k}\left[\boldsymbol{\alpha}_{f}\right]\right] + \mathbb{E}^{k}\left[\boldsymbol{\alpha}_{f}^{T}\right]\mathbb{E}^{k}\left[\mathbf{F}_{a_r}\right]^{T}\mathbb{E}^{k}\left[\mathbf{F}_{a_r}\right]\mathbb{E}^{k}\left[\boldsymbol{\alpha}_{f}\right] \\
&+ \mathrm{tr}\left[\mathbb{E}^{k}\left[\mathbf{F}_{a_r}\right]^{T}\mathbb{E}^{k}\left[\mathbf{F}_{a_r}\right]\mathrm{cov}^{k}\left[\boldsymbol{\alpha}_{f}\right]\right] + \mathrm{tr}\left[\mathbb{E}^{k}\left[\boldsymbol{A}_{f_{r_+}}\right]^{T}\mathbb{E}^{k}\left[\boldsymbol{A}_{f_{r_+}}\right]\mathrm{cov}^{k}\left[\mathbf{f}_{r_+}\right]\right]\Bigg),
\end{aligned}
\tag{F.78}
$$

where the last term is found using a similar derivation as that for the term $\mathbb{E}\left[\mathbf{h}^{T}\mathbf{F}^{T}\mathbf{F}\mathbf{h}\right]_{q^{k}(\mathbf{f},\mathbf{h})}$ in (F.64), and we make use of the identities $\mathbf{x}^{T}\mathbf{x} = \mathrm{tr}\left[\mathbf{x}\mathbf{x}^{T}\right]$ and $\mathbb{E}\left[\mathbf{x}\mathbf{x}^{T}\right] = \mathbb{E}\left[\mathbf{x}\right]\mathbb{E}\left[\mathbf{x}\right]^{T} + \mathrm{cov}\left[\mathbf{x}\right]$.

Finally, we may write

$$
\begin{aligned}
\mathbb{E}\left[\boldsymbol{\nu}_{r}^{T}\boldsymbol{\nu}_{r}\right]_{q(\mathbf{f},\mathbf{m}_f,\boldsymbol{\alpha}_f)} &= \|\left(\mathbf{I}_{+} - \mathbb{E}^{k}\left[\boldsymbol{A}_{f_{r_+}}\right]\right)\mathbb{E}^{k}\left[\mathbf{f}_{r_+}\right] - \left(\mathbf{I} - \mathbb{E}^{k}\left[\boldsymbol{A}_{m_r}\right]\right)\mathbb{E}^{k}\left[\mathbf{m}_{f_r}\right]\|^{2} \\
&+ \mathbb{E}^{k}\left[\mu_{r}\right]^{2}\mathbf{1}_{L_r}^{T}\,\mathrm{cov}^{k}\left[\boldsymbol{A}_{m_r}^{T}\right]\mathbf{1}_{L_r} + \left(\mathrm{cov}^{k}\left[\mu_{r}\right]\right)\mathbf{1}_{L_r}^{T}\left(\|\mathbf{I} - \mathbb{E}^{k}\left[\boldsymbol{A}_{m_r}\right]\|^{2} + \mathrm{cov}^{k}\left[\boldsymbol{A}_{m_r}^{T}\right]\right)\mathbf{1}_{L_r} \\
&- 2\mathbb{E}^{k}\left[\mu_{r}\right]\mathbf{1}_{L_r}^{T}\,\mathrm{cov}^{k}\left[\boldsymbol{A}_{m_r}^{T}, \boldsymbol{A}_{f_{r_+}}^{T}\right]\mathbb{E}^{k}\left[\mathbf{f}_{r_+}\right] \\
&+ \mathrm{tr}\left[\mathrm{cov}^{k}\left[\mathbf{f}_{r_+}\right]\right] - 2\mathbb{E}^{k}\left[\boldsymbol{\alpha}_{f_r}\right]^{T}\mathrm{cov}^{k}\left[\mathbf{F}_{a_r}^{T}, \mathbf{f}_{r_+}^{T}\right] + \Bigg(\mathrm{tr}\left[\mathrm{cov}^{k}\left[\mathbf{F}_{a_r}^{T}\right]\mathrm{cov}^{k}\left[\boldsymbol{\alpha}_{f}\right]\right] \\
&+ \mathrm{tr}\left[\mathbb{E}^{k}\left[\mathbf{F}_{a_r}\right]^{T}\mathbb{E}^{k}\left[\mathbf{F}_{a_r}\right]\mathrm{cov}^{k}\left[\boldsymbol{\alpha}_{f}\right]\right] + \mathrm{tr}\left[\mathbb{E}^{k}\left[\boldsymbol{A}_{f_{r_+}}\right]^{T}\mathbb{E}^{k}\left[\boldsymbol{A}_{f_{r_+}}\right]\mathrm{cov}^{k}\left[\mathbf{f}_{r_+}\right]\right]\Bigg).
\end{aligned}
\tag{F.79}
$$

The first term is just $\|\mathbb{E}^{k}\left[\boldsymbol{\nu}_{r}\right]\|^{2}$, where $\mathbb{E}^{k}\left[\boldsymbol{\nu}_{r}\right]$ is found simply by extracting the relevant pixels from $\mathbb{E}^{k}\left[\boldsymbol{\nu}\right] = \mathbb{E}^{k}\left[\left(\mathbf{I}_{+} - \boldsymbol{A}_{f}\right)\mathbf{f} - \left(\mathbf{I} - \boldsymbol{A}_{m}\right)\boldsymbol{\mu}\right] = \left(\mathbf{I}_{+} - \mathbb{E}^{k}\left[\boldsymbol{A}_{f}\right]\right)\mathbb{E}^{k}\left[\mathbf{f}\right] - \left(\mathbf{I} - \mathbb{E}^{k}\left[\boldsymbol{A}_{m}\right]\right)\mathbb{E}^{k}\left[\boldsymbol{\mu}\right]$.

F.8 Distribution Approximation for σ_u

In a similar manner to that for σ_ν, we have for σ_u:

$$
\log q^{k}\left(\sigma_{u}\right) \propto \mathbb{E}\left[\log\left(p\left(\mathbf{h}\,|\,\sigma_{u}\right)p\left(\sigma_{u}\right)\right)\right]_{q(\Theta_{\setminus\sigma_u})} \tag{F.80}
$$

$$
\propto \log(p\left(\sigma_{u}\right)) - \frac{1}{2}\sum_{r=1}^{R_h} L_r \log(\sigma_{u_r}^{2}) + \mathbb{E}\left[\sigma_{u_r}^{-2}\mathbf{u}_{r}^{T}\mathbf{u}_{r}\right]_{q(\mathbf{h})} \tag{F.81}
$$

$$
\Rightarrow q^{k}\left(\sigma_{u}\right) = \prod_{r=1}^{R_h}\mathcal{IG}\left(\sigma_{u_r}^{2}\,\middle|\,\alpha_{u_r} + \frac{L_r}{2}, \beta_{u_r} + \frac{1}{2}\mathbb{E}\left[\mathbf{u}_{r}^{T}\mathbf{u}_{r}\right]_{q(\mathbf{h})}\right) \tag{F.82}
$$

This is in the form of a product of inverse Gamma distributions over the variances in each region of the PSF. Hence to find $q^{k+1}\left(\sigma^2_{u_r}\right)$, we now just require the expectation

$$\mathbb{E}\left[\mathbf{u}_r^T\mathbf{u}_r\right]_{q(\mathbf{h})} = \mathbb{E}\left[\|\mathbf{C}_{r_+}\mathbf{h}_{r_+}\|^2\right]_{q(\mathbf{h})} \tag{F.83}$$

$$= \mathbb{E}^k\left[\mathbf{u}_r\right]^T\mathbb{E}^k\left[\mathbf{u}_r\right] + \mathrm{tr}\left[\mathbf{C}_{r_+}^T\mathbf{C}_{r_+}\,\mathrm{cov}^k\left[\mathbf{h}_{r_+}\right]\right] \tag{F.84}$$

$\mathbb{E}^k\left[\mathbf{u}_r\right]$ is found simply by extracting the relevant pixels from $\mathbb{E}^k\left[\mathbf{u}\right] = \mathbf{C}\mathbb{E}^k\left[\mathbf{h}\right]$.

Appendix G
Covariance of a Toeplitz Matrix

Here we elaborate on the definition $\text{cov}^k\left[\mathbf{X}^T\right] \triangleq \mathbb{E}^k\left[\mathbf{X}^T\mathbf{X} - \mathbb{E}^k[\mathbf{X}]^T\,\mathbb{E}^k[\mathbf{X}]\right]$. Considering a 1-D example for vector $\mathbf{x} = [x_1 \cdots x_m]^T$, from which a toeplitz convolution matrix $\mathbf{X}$ of dimension $m \times n$ is formed; the covariance of this matrix is defined with elements

$$\text{cov}\left[\mathbf{X}^T\right]_{[i,i']} \triangleq \sum_k \mathbb{E}\left[\Big(x(k-i) - \mathbb{E}\left[x(k-i)\right]\Big)\Big(x(k-i') - \mathbb{E}\left[x(k-i')\right]\Big)\right], \quad \forall i = 1\cdots n, i' = 1\cdots n, \tag{G.1}$$

$$= \sum_k \text{cov}\left[\mathbf{x}\right]_{[(i-k),(i-k')]} \tag{G.2}$$

where the summations are over properly defined elements x_i, according to the boundary conditions used. For the 2-D case where BTTB convolution matrices are employed, the corresponding definition is

$$\text{cov}\left[\mathbf{H}^T\right]_{[m(j-1)+i,m(j'-1)+i']} \tag{G.3}$$

$$\triangleq \sum_{k,l} \mathbb{E}\left[\Big(h(k-i,l-j) - \mathbb{E}\left[h(k-i,l-j)\right]\Big)\Big(h(k-i',l-j') - \mathbb{E}\left[h(k-i',l-j')\right]\Big)\right],$$

$$\forall i, i' = 1\cdots m,\ j, j' = 1\cdots n, \tag{G.4}$$

$$= \sum_{k,l} \text{cov}\left[\mathbf{h}\right]_{[m(i-k)+(j-l),m(i-k')+(j-l')]} \tag{G.5}$$

again for appropriately defined elements of the 2-D PSF image $h(k,l)$, according to the boundary conditions. Observe that if circulant boundary conditions are used, and the vectors f and h are of equal size, this expression reduces to $\text{cov}\left[\mathbf{H}^T\right] = mn\,\text{cov}\left[\mathbf{h}\right]$.

Appendix H
Video clips

The iterations from the VB experiments and the samples from the MCMC experiments have been made into short movie clips that make it very straightforward to compare the progress of the algorithms, and to see how the solutions evolve. It is informative to see the visual effects to gain some insight into the way the algorithms behave. These video clips are made available on the accompanying DVD, in uncompressed AVI format. The following table lists the figures that correspond to each clip.

Table H.1: *Video clips of the experiments*

Filename	Figure	Page no.	Details
Ch10clip1	9.3(b)	164	
Ch10clip2	9.3(d)	164	
Ch10clip3	9.5(b)	166	
Ch10clip4	9.5(d)	166	
Ch10clip5	9.7(a)	167	
Ch10clip6	9.7(b)	167	
Ch10clip7	9.8(a)	169	
Ch10clip8	9.8(b)	169	
Ch10clip9	9.8(c)	169	
Ch10clip10	9.8(d)	169	
Ch10clip11	9.12(a)	175	
Ch10clip12	9.12(c)	175	
Ch10clip13	9.12(e)	175	
Ch10clip14	9.13(a)	176	
Ch10clip15	9.13(c)	176	

Filename	Figure	Page no.	Details
Ch10clip16	9.13(e)	176	
Ch10clip17	9.14(c)	179	
Ch10clip18	9.14(d)	179	
Ch10clip19	9.14(e)	179	
Ch10clip20	9.14(f)	179	
Ch10clip21	9.14(g)	179	
Ch10clip22	9.14(h)	179	
Ch10clip23	9.14(i)	179	
Ch10clip24	9.14(j)	179	
Ch10clip25	9.14(k)	179	
Ch10clip26	9.14(l)	179	
Ch10clip27	9.15(c)	180	
Ch10clip28	9.15(d)	180	
Ch10clip29	9.15(e)	180	
Ch10clip30	9.15(f)	180	
Ch10clip31	9.15(g)	180	
Ch10clip32	9.15(h)	180	
Ch10clip33	9.15(i)	180	
Ch10clip34	9.15(j)	180	
Ch10clip35	9.15(k)	180	
Ch10clip36	9.15(l)	180	
Ch10clip37	9.16(c)	181	
Ch10clip38	9.16(d)	181	
Ch10clip39	9.16(e)	181	
Ch10clip40	9.16(f)	181	
Ch10clip41	9.16(g)	181	
Ch10clip42	9.16(h)	181	

Filename	Figure	Page no.	Details
Ch10clip43	9.16(i)	181	
Ch10clip44	9.16(j)	181	
Ch10clip45	9.16(k)	181	
Ch10clip46	9.16(l)	181	
Ch10clip47	9.17(c)	182	
Ch10clip48	9.17(d)	182	
Ch10clip49	9.17(e)	182	
Ch10clip50	9.17(f)	182	
Ch10clip51	9.17(g)	182	
Ch10clip52	9.17(h)	182	
Ch10clip53	9.17(i)	182	
Ch10clip54	9.17(j)	182	
Ch10clip55	9.17(k)	182	
Ch10clip56	9.17(l)	182	
Ch11clip1	10.1(a)	188	
Ch11clip2	10.1(b)	188	
Ch11clip3	10.3(b)	195	Exp. 1
Ch11clip4	10.3(e)	195	Exp. 2
Ch11clip5	10.3(i)	195	Exp. 2
Ch11clip6	10.3(h)	195	Exp. 3
Ch11clip7	10.3(i)	195	Exp. 3
Ch11clip8	10.3(k)	195	Exp. 4
Ch11clip9			Exp. 5 (no figure)
Ch11clip10	10.6(b)	199	Exp. 6
Ch11clip11	10.6(e)	199	Exp. 7
Ch11clip12	10.6(h)	199	Exp. 8
Ch11clip13	10.6(k)	199	Exp. 9

Part IV

Abbreviations, Nomenclature & References

Acronyms and abbreviations

ADC	analogue to digital converter. 224
AM	alternating minimization. 63
AR-AR	autoregressive-autoregressive. 86
ARMA	autoregressive moving average. 13
AR	autoregressive. ii
BCCB	block circulant with circulant blocks. 30
BC	boundary condition. 31
BD	blind deconvolution. 37
BIBO	bounded input bounded output. 234
BID	blind image deconvolution. ii
BIR	blind image restoration. 92
BSAR	block-stationary AR. 82
BSNR	blurred-image SNR. 122
BSR	blind superresolution. 204
BSS	blind source separation. 72
BTTB	block Toeplitz with Toeplitz blocks. 19
CAR	conditional autoregression. 53
CCD	charge coupled device. 52
CDF	cumulative distribution function.
CGLS	conjugate gradients least sqaures method. 33
CGMRF	compound GMRF. 56
CGM	compound Gauss-Markov. 232
CG	conjugate gradients method. 32
CLS	constrained least squares filter. 28

CMOS complementary metal oxide semiconductor. 223

CM conditional Markov. 228

CoC circle of confusion. 51

CPU central processing unit. 198

CT computed tomography. 7

DAG directed acyclic graph. 45

DCT discrete cosine transform. 52

DFT discrete Fourier transform. 31

DoF depth of field. 216

DPC discrete Picard condition. 29

DSG doubly stochastic Gaussian. 232

EM expectation maximization algorithm. 65

FFT fast Fourier transform. 135

FIR finite impulse response. 47

FoE fields of experts. 202

FT Fourier transform. 50

GCD greatest common divisor. 248

GCV generalised cross-validation. 29

GEM generalised expectation maximization algorithm. 67

GGMRF generalised Gaussian MRF. 56

GMRF Gaussian Markov random field. 56

GPGPU general purpose graphics processing unit. 200

GPR ground penetrating radar. 8

GPU graphics processing unit. 204

GSM Gaussian scale mixture. 202

HMM Hidden Markov Model. 232

HMRF Huber MRF. 232

HOS higher order statistics. 243

HR	high resolution. 204
HST	Hubble space telescope. 7
HVS	human visual system. 30
IBD	iterative blind deconvolution. 244
ICM	iterated conditional modes. 63
IDFT	inverse discrete Fourier transform.
IG	inverse-Gamma (distribution). 44
iid	independent and identically distributed. 88
ISNR	improvement in signal-to-noise ratio. 34
IVB	iterative variational Bayesian. 71
KL	Kullback-Leibler (distance). 67
LSI	linear spatially-invariant. ii
LSV	linear spatially-varying. 17
LS	least squares filter. 27
MAP	maximum a posteriori. 61
MA	moving average. 54
MBD	multichannel blind deconvolution. 204
MCMC	Markov chain Monte Carlo. 14
MC	Markov chain. 227
MEM	maximum entropy methods. 239
MIMO	multiple input multiple output. 204
ML	maximum likelihood. 13
MMAP	maximum marginalised *a posteriori*. ii
MMSE	minimum mean-squared error. 64
MRF	Markov random field. x
MSE	mean squared error. 30
MTF	modulation transfer function. 51
NAS-RIF	nonnegativity and support constraints with recursive image filtering. 135

OCR	optical character recognition.	8
OTF	optical transfer function.	51
PCA	principal component analysis.	247
PCG	preconditioned conjugate gradients method.	34
PDE	partial differential equation.	63
PDF	probability density function.	23
PIF	point-influence function.	xii
POCS	projection onto convex sets.	35
PSF	point-spread function.	xii
RCTLS	regularized constrained total least squares.	247
RGB	red, green and blue.	221
RHS	right hand side.	173
SAR	simultaneous autoregression.	53
SA	simulated annealing.	188
SD	steepest descent method.	32
SIMO	single input multiple output.	204
SI	spatially invariant.	18
SNR	signal-to-noise ratio.	4
SPECT	single photon emission computed tomography.	7
SR	superresolution.	29
SSIM	structural similarity measure.	35
SVD	singular value decomposition.	141
SV	spatially varying.	18
TLS	total least squares.	
TV	total variation.	57
VB	variational Bayesian.	ii
VQ	vector quantizer.	247
WGN	white Gaussian noise.	41

w.r.t with respect to. 27

ZT Z-transform. 243

Nomenclature

General Symbols

$\boldsymbol{x}$	Vectors denoted by bold lowercase letters
$\boldsymbol{A}$	Matrices denoted by bold uppercase letters
$[\boldsymbol{A}]_{i,j}$	element at row i, column j of matrix
$\boldsymbol{x}^T, \boldsymbol{A}^T$	Transposed vector or matrix
$\boldsymbol{A}^{-1}$	Inverse of a square matrix
$\boldsymbol{A}^+$	Psuedo-inverse
$\backslash$	Gaussian elimination operator (MATLAB style)
$\|\boldsymbol{x}\|^2$	L_2 norm of a vector
$\det \boldsymbol{A}$	Matrix determinant
$\|\mathcal{S}\|$	Set size
$\mathrm{tr}\,[\boldsymbol{A}]$	Trace of a matrix
vec	Vectorisation (Lexicographic ordering)
diag	Convert vector to diagonal matrix or vice versa
$\hat{\boldsymbol{x}}$	Estimate of a vector
$\otimes$	Kronecker product
$\mathbf{I}_L$	Identity matrix of size $L \times L$
$\mathbf{0}_L$	vector of L zeros
$\mathbf{1}_L$	vector of L ones
$\star$	Convolution
$\cdot$	Scalar multiplication or vector inner product
∇	Gradient
$\nabla\cdot$	Divergence
$\triangleq$	Is defined as
$\sim$	Is distributed as
$\leftarrow$	Variable is sampled from the distribution on the right
$\rightarrow$	Tends to in the limit
$\in$	Member of
$\subset$	Subset

θ	Model parameter
$\boldsymbol{\Theta}$	Set of parameters
$\boldsymbol{\Theta}_{\setminus\theta}$	Set of parameters, excluding variable θ
$\mathbb{R}$	Set of real numbers
$f\vert_x$	Function f evaluated at a fixed point x
$p(\mathbf{x})$	PDF of $\mathbf{x}$
$p(\mathbf{x} \mid \mathbf{y})$	Conditional PDF of $\mathbf{x}$ given $\mathbf{y}$
$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$	Gaussian distribution in $\mathbf{x}$ with mean $\boldsymbol{\mu}$ and Covariance $\boldsymbol{\Sigma}$
$\mathcal{N}^{+}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$	Rectified Gaussian distribution in $\mathbf{x}$ with mean $\boldsymbol{\mu}$ and Covariance $\boldsymbol{\Sigma}$
$\mathcal{GA}(\theta \mid \alpha, \beta)$	Gamma distribution in θ with parameters α, β
$\mathcal{IG}(\theta \mid \alpha, \beta)$	Inverse Gamma distribution in θ with parameters α, β
$C_{\mathrm{KL}}(\cdot \parallel \cdot)$	Kullback-Leibler (KL) divergence
$\mathrm{Var}[\theta]$	Variance
$\mathrm{cov}[\mathbf{x}]$	Covariance
$\mathbb{E}[\mathbf{x}]_{p(\mathbf{x})}$	Expected value of $\mathbf{x}$ w.r.t distribution $p(\mathbf{x})$
$\mathbb{E}^{k}[\mathbf{x}]$	Expected value of $\mathbf{x}$ at iteration k (for VB approach)
$\mathrm{cov}^{k}[\mathbf{x}]$	covariance of $\mathbf{x}$ at iteration k (for VB approach)
$q^{k}(\theta)$	Approximate distribution in θ at iteration k (for VB approach)

References

[1] E. H. Adelson and J. R. Bergen, "The plenoptic function and the elements of early vision," in *Computation Models of Visual Processing*, M. Landy and J. Movshon, Eds. MIT Press, 1991, pp. 3–20. [Online]. Available: http://web.mit.edu/persci/people/adelson/pub_pdfs/elements91.pdf

[2] E. H. Adelson and J. Y. Wang, "Single lens stereo with a plenoptic cameras," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 99–106, 1992.

[3] F. Aghdasi and R. Ward, "Reduction of boundary artifacts in image restoration," *Image Processing, IEEE Transactions on*, vol. 5, no. 4, pp. 611–618, 1996.

[4] A. Agrawal and R. Raskar, "Resolving objects at higher resolution from a single motion-blurred image," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, 2007, pp. 1–8.

[5] H. C. Andrews and B. R. Hunt, *Digital Image Restoration*. Englewood Cliffs, New Jersey: Prentice-Hall, 1977.

[6] C. Andrieu, N. de Freitras, A. Doucet, and M. Jordan, "An introduction to MCMC for machine learning," *Machine Learning*, vol. 50, pp. 5–43, 2003.

[7] G. Ayers and J. Dainty, "Iterative blind deconvolution method and its applications," *Optics Letters*, vol. 13, no. 7, pp. 547–549, 1988.

[8] S. D. Babacan, S. D. Babacan, R. Molina, and A. K. Katsaggelos, "Parameter estimation in tv image restoration using variational distribution approximation," *Image Processing, IEEE Transactions on*, vol. 17, no. 3, pp. 326–339, 2008.

[9] S. D. Babacan, R. Molina, and A. K. Katsaggelos, "Parameter estimation in total variation blind deconvolution," in *European Signal Processing Conference (EUSIPCO)*. Lausanne (Switzerland), September 2008. [Online]. Available: http://decsai.ugr.es/vip/files/conferences/1569105302.pdf

[10] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, pp. 1167–1183, 2002.

[11] M. R. Banham, N. P. Galatsanos, H. L. Gonzalez, and A. K. Katsaggelos, "Multichannel restoration of single channel images using a wavelet-based subband decomposition," *IEEE Transactions on Image Processing*, vol. 3, no. 6, pp. 821–833, 1994. [Online]. Available: http://dx.doi.org/10.1109/83.336250

[12] M. Banham and A. Katsaggelos, "Digital image restoration," *IEEE Signal Processing Magazine*, vol. 14, no. 2, pp. 24–41, 1997.

[13] R. Barnard, V. P. Pauca, T. C. Torgersen, R. J. Plemmons, S. Prasad, J. van der Gracht, J. Nagy, J. Chung, G. Behrmann, S. Mathews, and M. Mirotznik, "High-resolution iris image reconstruction from low-resolution imagery," in *Proc. SPIE Int. Soc. Opt. Eng.*, vol. 6313. San Diego, CA, USA: SPIE, Aug. 2006, pp. 63 130D–13. [Online]. Available: http://link.aip.org/link/?PSI/6313/63130D/1

[14] M. J. Beal, "Variational algorithms for approximate bayesian inference," Ph.D. dissertation, Gatsby Computational Neuroscience Unit, University College London, 2003.

[15] M. Ben-Ezra and S. K. Nayar, "Motion-based motion deblurring," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 6, pp. 689–698, 2004.

[16] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*. New York, Springer Verlag, 1985, ch. 3 and 4.

[17] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*. John Wiley & Sons, 1994.

[18] M. Bertero and P. Boccacci, *Introduction to inverse problems in imaging*, 1st ed. Institute of Physics Publishing, 1998.

[19] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 36, no. 2, pp. 192–236, 1974. [Online]. Available: http://links.jstor.org/sici?sici=0035-9246%281974%2936%3A2%3C192%3ASIATSA%3E2.0.CO%3B2-3

[20] ——, "On the statistical analysis of dirty pictures," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 48, no. 3, pp. 259–302, 1986. [Online]. Available: http://links.jstor.org/sici?sici=0035-9246%281986%2948%3A3%3C259%3AOTSAOD%3E2.0.CO%3B2-%23

[21] J. Biemond, R. L. Lagendijk, and R. M. Mersereau, "Iterative methods for image deblurring," *Proceedings of the IEEE*, vol. 78, no. 5, pp. 856–883, May 1990.

[22] J. M. Bioucas-Dias, M. A. T. Figueiredo, and J. P. Oliveira, "Total variation-based image deconvolution: A majorization-minimization approach," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, May 2006.

[23] T. E. Bishop, "Blind restoration of blurred photographs," Fourth year undergraduate project in Group F, Department of Engineering, University of Cambridge, May 2004. [Online]. Available: http://www.tombishop.net

[24] T. E. Bishop and J. R. Hopgood, "Blind image restoration using a block-stationary signal models," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 2, May 2006, pp. II–853–II–856.

[25] T. E. Bishop, S. D. Babacan, B. Amizic, T. Chan, R. Molina, and A. K. Katsaggelos, "Blind image deconvolution: problem formulation and existing approaches," in *Blind image deconvolution: Theory and Applications*, P. Campisi and K. Egiazarian, Eds. CRC, 2007, ch. 1.

[26] T. E. Bishop, R. Molina, and J. R. Hopgood, "Nonstationary blind image restoration using variational methods," in *IEEE International Conference on Image Processing (ICIP)*, 2007.

[27] ——, "Blind restoration of blurred photographs via AR modelling and MCMC," in *IEEE International Conference on Image Processing (ICIP)*, 2008.

[28] S. Borman and R. L. Stevenson, "Super-resolution from image sequences – A review," in *Proceedings of the 1998 Midwest Symposium on Circuits and Systems*. Notre Dame, IN, USA: IEEE, aug 1998, pp. 374–378. [Online]. Available: http://www.seanborman.com/publications/mwscas98.pdf

[29] M. Born and E. Wolf, *Principles of Optics*. Pergamon Press, 1959.

[30] N. Bose and K. Boo, "Asymptotic eigenvalue distribution of block-Toeplitz matrices," *IEEE Transactions on Information Theory*, vol. 44, no. 2, pp. 858 – 861, 1998. [Online]. Available: http://dx.doi.org/10.1109/18.661535

[31] ——, "High-resolution image reconstruction with multisensors," *International Journal of Imaging Systems and Technology*, vol. 9, no. 4, pp. 294–304, 1998. [Online]. Available: http://www3.interscience.wiley.com/cgi-bin/abstract/78501/ABSTRACT

[32] M. Boudaoud and L. Chaparro, "Nonstationary composite modeling of images," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 19, no. 1, pp. 112–117, 1989.

[33] C. A. Bouman, "Markov random fields and stochastic image models," Tutorial Presentation, IEEE International Conference on Image Processing, Washington D.C., Oct 1995, available online at http://dynamo.ecn.purdue.edu/~bouman/publications/Index-Tutorials.html. [Online]. Available: http://dynamo.ecn.purdue.edu/~bouman/publications/Index-Tutorials.html

[34] C. A. Bouman and K. Sauer, "Generalized gaussian image model for edge-preserving map estimation," *IEEE Transactions on Image Processing*, vol. 2, no. 3, pp. 296–310, 1993.

[35] C. A. Bouman and M. Shapiro, "Multiscale random field model for Bayesian image segmentation," *IEEE Transactions on Image Processing*, vol. 3, no. 2, pp. 162–177, 1994.

[36] D. Brook, "On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems," *Biometrika*, vol. 51, no. 3/4, pp. 481–483, Dec 1964, biometric Laboratory, University College, London. [Online]. Available: http://links.jstor.org/sici?sici=0006-3444%28196412%2951%3A3%2F4%3C481%3AOTDBTC%3E2.0.CO%3B2-S

[37] D. Calvetti and E. Somersalo, "Statistical elimination of boundary artefacts in image deblurring," *Inverse Probl*, vol. 21, no. 5, pp. 1697–1714, 2005. [Online]. Available: http://www.scopus.com/scopus/inward/record.url?eid=2-s2.0-25444490175&partnerID=40&rel=R5.6.0

[38] D. Calvetti, J. Kaipio, and E. Somersalo, "Aristotelian prior boundary conditions," *International Journal of Mathematics and Computer Science*, vol. 1, pp. 63–81, 2006.

[39] D. Calvetti and E. Somersalo, "Bayesian image deblurring and boundary effects," *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 5910, no. 1, pp. 59 100X–9, Aug. 2005. [Online]. Available: http://link.aip.org/link/?PSI/5910/59100X/1

[40] P. Campisi and G. Scarano, "A multiresolution approach for texture synthesis using the circular harmonic functions," *IEEE Trans. Image Process.*, vol. 11, no. 1, pp. 37–51, Jan 2002.

[41] P. Campisi, S. Colonnese, G. Panci, and G. Scarano, "Multichannel bussgang algorithm for blind restoration of natural images," in *Image Processing, 2003. Proceedings. 2003 International Conference on*, vol. 3, 2003, pp. II–985–988.

[42] M. Cannon, "Blind deconvolution of spatially invariant image blurs with phase," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 1, pp. 58–63, 1976.

[43] Canon Inc., *EF Lens Work III*, 8th ed. Canon Inc., Sep 2006. [Online]. Available: http://www.canon-europe.com/Support/Documents/digital_slr_educational_tools/en/ef_lens_work_iii_en.asp

[44] W. Cathey and E. Dowski, "New paradigm for imaging systems," *Applied Optics*, vol. 41, no. 29, pp. 6080–6092, 2002.

[45] A. Chambolle and P.-L. Lions, "Image recovery via total variation minimization and related problems," *Numerische Mathematik*, vol. 76, no. 2, pp. 167 – 88, 1997. [Online]. Available: http://dx.doi.org/10.1007/s002110050258

[46] R. H. Chan, J. G. Nagy, and R. J. Plemmons, "FFT-based preconditioners for Toeplitz-block least squares problems," *SIAM Journal on Numerical Analysis*, vol. 30, no. 6, pp. 1740–1768, 1993.

[47] R. Chan and M. K. Ng, "Conjugate gradient methods for Toeplitz systems," *SIAM Review*, vol. 38, no. 3, pp. 427–482, 1996.

[48] R. Chan, T. Chan, M. Ng, and A. Yip, "Cosine transform preconditioners for high resolution image reconstruction," *Linear Algebra and Its Applications*, vol. 316, no. 1-3, pp. 89–104, 2000.

[49] T. C. Chan and J. Shen, *Image Processing And Analysis: Variational, Pde, Wavelet, And Stochastic Methods*. SIAM, 2005.

[50] T. F. Chan and C.-K. Wong, "Total variation blind deconvolution," *Image Processing, IEEE Transactions on*, vol. 7, no. 3, pp. 370–375, 1998.

[51] T. F. Chan, G. H. Golub, and P. Mulet, "A nonlinear primal-dual method for total variation-based image restoration," *SIAM J. Sci. Comput.*, vol. 20, no. 6, pp. 1964–1977, November 1999. [Online]. Available: http://dx.doi.org/10.1137/S1064827596299767

[52] T. F. Chan, S. Osher, and J. Shen, "Digital tv filter and nonlinear denoising," *IEEE Transactions on Image Processing*, vol. 10, no. 2, pp. 231–241, 2001. [Online]. Available: http://www.scopus.com/scopus/inward/record.url?eid=2-s2.0-0035250066&partner=40&rel=R4.5.0

[53] W.-S. Chan, E. Lam, M. Ng, and G. Mak, "Super-resolution reconstruction in a computational compound-eye imaging system," *Multidimensional Systems and Signal Processing*, vol. 18, no. 2, pp. 83–101, Sep. 2007. [Online]. Available: http://dx.doi.org/10.1007/s11045-007-0022-3

[54] M. M. Chang, A. M. Tekalp, and A. T. Erdem, "Blur identification using the bispectrum," *IEEE Transactions on Signal Processing*, vol. 39, no. 10, pp. 2323–2325, 1991. [Online]. Available: http://dx.doi.org/10.1109/78.91207

[55] R. Chellappa and R. L. Kashyap, "Texture synthesis using 2-D noncausal autoregressive models." *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-33, no. 1, pp. 194–203, 1985.

[56] R. Chellappa and R. Kashyap, "Digital image restoration using spatial interaction models." *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-30, no. 3, pp. 461–472, 1982.

[57] L. Chen and K.-H. Yap, "A soft double regularization approach to parametric blind image deconvolution," *Image Processing, IEEE Transactions on*, vol. 14, no. 5, pp. 624–633, 2005.

[58] B. A. Chipman and B. D. Jeffs, "Blind multiframe point source image restoration using MAP estimation," *Conference Record of the Asilomar Conference on Signals, Systems and Computers*, vol. 2, pp. 1267–1271, 1999.

[59] S. Dai, M. Yang, Y. Wu, and A. K. Katsaggelos, "Tracking motion-blurred targets in video," in *International Conference on Image Processing (ICIP'06)*, Atlanta, GA, October 2006.

[60] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977. [Online]. Available: http://links.jstor.org/sici?sici=0035-9246%281977%2939%3A1%3C1%3AMLFIDV%3E2.0.CO%3B2-Z

[61] H. Derin and H. Elliott, "Modelling and segmentation of noisy and textured images using gibbs random fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-9, no. 1, pp. 39–55, Jan 1987.

[62] A. Doucet and X. Wang, "Monte carlo methods for signal processing: a review in the statistical signal processing context," *Signal Processing Magazine, IEEE*, vol. 22, no. 6, pp. 152–170, 2005.

[63] S. N. Efstratiadis and A. K. Katsaggelos, "Adaptive iterative image restoration with reduced computational load," *Optical Engineering*, vol. 29, pp. 1458–1468, 1990.

[64] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *Image Processing, IEEE Transactions on*, vol. 15, no. 12, pp. 3736–3745, 2006. [Online]. Available: http://www.cs.technion.ac.il/~elad/publications/journals/2005/KSVD_Denoising_IEEE_TIP.pdf

[65] M. Elad and D. Datsenko, "Example-based regularization deployed to super-resolution reconstruction of a single image," *The Computer Journal*, vol. 50, no. 4, pp. 1–16, April 2007.

[66] S. Farsiu, D. Robinson, P. Milanfar, and M. Elad, "Advances and challenges in super-resolution," *International Journal of Imaging Systems and Technology*, vol. 14, no. 2, pp. 47–57, 2004. [Online]. Available: http://www.cse.ucsc.edu/~milanfar/SR-challengesIJIST.pdf

[67] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W. T. Freeman, "Removing camera shake from a single photograph," in *ACM Transactions on Graphics (Proc. SIGGRAPH 2006)*, vol. 25, University of Toronto, 2006, pp. 787–794. [Online]. Available: http://www.scopus.com/scopus/inward/record.url?eid=2-s2.0-33749249573&partnerID=40&rel=R5.5.0

[68] B. Frieden, "Restoring with maximum likelihood and maximum entropy." *Journal of the Optical Society of America*, vol. 62, no. 4, pp. 511–518, 1972.

[69] N. P. Galatsanos, M. Wernick, and A. K. Katsaggelos, "Multi-channel image recovery," in *Handbook of Image and Video Processing*, A. Bovik, Ed. Academic Press, 2000, ch. 3.7, pp. 161–174.

[70] N. P. Galatsanos and A. K. Katsaggelos, "Methods for choosing the regularization parameter and estimating the noise variance in image restoration and their relation," *IEEE Transactions on Image Processing*, vol. 1, no. 3, pp. 322–336, 1992. [Online]. Available: http://dx.doi.org/10.1109/83.148606

[71] N. P. Galatsanos, V. Z. Mesarović, R. Molina, and A. K. Katsaggelos, "Hierarchical bayesian image restoration from partially known blurs," *IEEE Transactions on Image Processing*, vol. 9, no. 10, pp. 1784–1797, 2000. [Online]. Available: http://ivpl.ece.northwestern.edu/Publications/Journals/2000/IEEETransIP00b.pdf

[72] N. P. Galatsanos, V. Mesarović, R. Molina, A. Katsaggelos, and J. Mateos, "Hyperparameter estimation in image restoration problems with partially known blurs," *Optical Engineering*, vol. 41, no. 8, pp. 1845–1854, 2002. [Online]. Available: http://ivpl.ece.northwestern.edu/Publications/Journals/2002/cvips.pdf

[73] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, 2nd ed. Chapman & Hall, 2004.

[74] D. Geman and G. Reynolds, "Constrained restoration and the recovery of discontinuities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 3, pp. 367–383, 1992.

[75] D. Geman and C. Yang, "Nonlinear image recovery with half-quadratic regularization," *IEEE Transactions on Image Processing*, vol. 4, no. 7, pp. 932–946, 1995.

[76] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-6, no. 6, pp. 721–741, 1984.

[77] G. Giannakis and J. Heath, R.W., "Blind identification of multichannel FIR blurs and perfect image restoration," *Image Processing, IEEE Transactions on*, vol. 9, no. 11, pp. 1877–1896, 2000.

[78] W. Gilks, S. Richardson, and D. Spiegelhalter, Eds., *Markov Chain Monte Carlo in Practice: Interdisciplinary Statistics*. Chapman & Hall/CRC, 1995.

[79] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 1st ed. Addison Wesley, 1992. [Online]. Available: http://www.imageprocessingplace.com/

[80] I. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 6th ed., A. Jeffrey and D. Zwillinger, Eds. Academic Press, jul 2000.

[81] R. M. Gray, "Toeplitz and circulant matrices: a review," available online at http://www-ee.stanford.edu/~gray/toeplitz.html, 2002. [Online]. Available: http://www-ee.stanford.edu/~gray/toeplitz.html

[82] ——, "On the asymptotic eigenvalue distribution of Toeplitz matrices." *IEEE Transactions on Information Theory*, vol. IT-18, no. 6, pp. 725–730, 1972. [Online]. Available: http://dx.doi.org/10.1109/TIT.1972.1054924

[83] J. Guerrero-Colon, L. Mancera, and J. Portilla, "Image restoration using space-variant gaussian scale mixtures in overcomplete pyramids," *Image Processing, IEEE Transactions on*, vol. 17, no. 1, pp. 27–41, 2008.

[84] Y. P. Guo, H. P. Lee, and C. L. Teo, "Blind restoration of images degraded by space-variant blurs using iterative algorithms for both blur identification and image restoration," *Image and Vision Computing*, vol. 15, no. 5, pp. 399–410, May 1997. [Online]. Available: http://www.sciencedirect.com/science/article/B6V09-3SNV46R-B/2/f0adb733256285e4c1573a0bf5d3407a

[85] A. Hamza, H. Krim, and G. Unal, "Unifying probabilistic and variational estimation," *IEEE Signal Processing Magazine*, vol. 19, no. 5, pp. 37–47, 2002.

[86] P. Hansen, "Deconvolution and regularization with toeplitz matrices," *Numerical Algorithms*, vol. 29, no. 4, pp. 323–378, 2002.

[87] P. C. Hansen, "The discrete Picard condition for discrete ill-posed problems," *BIT*, vol. 30, no. 4, pp. 658–672, 1990.

[88] ——, "Analysis of discrete ill-posed problems by means of the L-curve," *SIAM Review*, vol. 34, no. 4, pp. 561–580, Dec 1992. [Online]. Available: http://links.jstor.org/sici?sici=0036-1445%28199212%2934%3A4%3C561%3AAODIPB%3E2.0.CO%3B2-J

[89] G. Harikumar and Y. Bresler, "Exact image deconvolution from multiple FIR blurs," *IEEE Transactions on Image Processing*, vol. 8, no. 6, pp. 846–862, 1999.

[90] ——, "Perfect blind restoration of images blurred by multiple filters:theory and efficient algorithms," *Image Processing, IEEE Transactions on*, vol. 8, no. 2, pp. 202–219, Feb 1999.

[91] W. K. Hastings, "Monte carlo sampling methods using markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970. [Online]. Available: http://dx.doi.org/10.2307/2334940

[92] J. R. Hopgood, "Nonstationary signal processing with application to reverberation cancellation in acoustic environments," Ph.D. dissertation, Signal Processing Laboratory, Department of Engineering, University of Cambridge, November 2000. [Online]. Available: http://www.james-hopgood.net/

[93] J. R. Hopgood and P. J. W. Rayner, "Blind single channel deconvolution using nonstationary signal processing," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 7, pp. 476–488, September 2003. [Online]. Available: http://www.james-hopgood.net/Users/jrh/?doc=Publications&type=all

[94] B. R. Hunt, "Matrix theory proof of the discrete convolution theorem," *IEEE Trans. Audio and Electroacoustics*, vol. AU-19, no. 4, pp. 285–288, Dec 1971.

[95] ——, "Application of constrained least squares estimation to image restoration by digital computer." *IEEE Transactions on Computers*, vol. C-22, no. 9, pp. 805–812, 1973.

[96] A. K. Jain, *Fundamentals of Digital Image Processing*, 1st ed. New Jersey: Prentice Hall, 1989.

[97] A. Jalobeanu, J. Zerubia, and L. Blanc-Féraud, "Bayesian estimation fo blur and noise in remote sensing imaging," in *Blind image deconvolution: Theory and Applications*, P. Campisi and K. Egiazarian, Eds. CRC, 2007, ch. 6.

[98] F.-C. Jeng and J. W. Woods, "Simulated annealing in compound Gaussian random fields," *IEEE Transactions on Information Theory*, vol. 36, no. 1, pp. 94–107, 1990.

[99] ——, "Compound Gauss-Markov random fields for image estimation," *IEEE Transactions on Signal Processing*, vol. 39, no. 3, pp. 683–697, 1991.

[100] M. I. Jordan, Z. Ghahramani, T. S. Jaakola, and L. K. Saul, "An introduction to variational methods for graphical models," in *Learning in Graphical Models*. MIT Press, 1998, pp. 105–162.

[101] J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, ser. Applied Mathematical Sciences. Springer, 2005, vol. 160.

[102] M. G. Kang and A. K. Katsaggelos, "General choice of the regularization functional in regularized image restoration," *IEEE Trans. Image Process.*, vol. 4, no. 5, pp. 594–602, 1995.

[103] R. L. Kashyap and R. Chellappa, "Estimation and choice of neighbors in spatial-interaction models of images," *Information Theory, IEEE Transactions on*, vol. 29, no. 1, pp. 60–72, 1983.

[104] R. Kass and A. E. Raftery, "Bayes factors," *Journal of the American Statistical Association*, vol. 90, pp. 773–795, 1995.

[105] A. K. Katsaggelos, "Iterative image restoration algorithms," Ph.D. dissertation, Georgia Institute of Technology, School of Electrical Engineering, August 1985.

[106] A. K. Katsaggelos, Ed., *Digital Image Restoration*. New York, Springer–Verlag, 1991.

[107] A. K. Katsaggelos and K. T. Lay, "Maximum likelihood identification and restoration of images using the expectation-maximization algorithm," in *Digital Image Restoration*, A. K. Katsaggelos, Ed. Springer-Verlag, 1991.

[108] A. K. Katsaggelos and K. Lay, "Maximum likelihood blur identification and image restoration using the EM algorithm," *IEEE Transactions on Signal Processing*, vol. 39, no. 3, pp. 729–733, 1991.

[109] A. K. Katsaggelos, J. Biemond, R. W. Schafer, and R. M. Mersereau, "A regularized iterative image restoration algorithm," *IEEE Transactions on Signal Processing*, vol. 39, no. 4, pp. 914–929, April 1991.

[110] A. Katsaggelos, J. Biemond, R. Mersereau, and R. Schafer, "Nonstationary iterative image restoration." in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 1985, pp. 696–699.

[111] A. Katsaggelos, K. Lay, and N. Galatsanos, "A general framework for frequency domain multi-channel signal processing," *IEEE Transactions on Image Processing*, vol. 2, no. 3, pp. 417–420, 1993. [Online]. Available: http://dx.doi.org/10.1109/83.236528

[112] A. Katsaggelos, R. Molina, and J. Mateos, *Super resolution of images and video*, ser. Synthesis Lectures on Image, Video, and Multimedia Processing. Morgan & Claypool, 2007. [Online]. Available: http://decsai.ugr.es/vip/files/books/SRbook.html

[113] A. Kokaram, "A statistical framework for picture reconstruction using 2d ar models," *Image and Vision Computing*, vol. 22, no. 2, pp. 165–171, 2004.

[114] J. Krist, "Simulation of HST PSFs using Tiny Tim," in *Astronomical Data Analysis Software and Systems IV*, R. A. Shaw, H. E. Payne, and J. J. E. Hayes, Eds. San Francisco, USA: Astronomical Society of the Pacific, 1995, pp. 349–353.

[115] D. Kundur and D. Hatzinakos, "Blind image deconvolution revisited," *IEEE Signal Processing Magazine*, vol. 13, no. 6, pp. 61–63, November 1996.

[116] ——, "Blind image deconvolution," *IEEE Signal Processing Magazine*, vol. 13, no. 3, pp. 43–64, May 1996.

[117] ——, "A novel blind deconvolution scheme for image restoration using recursive filtering," *IEEE Transactions on Signal Processing*, vol. 46, no. 2, pp. 375–390, 1998. [Online]. Available: citeseer.nj.nec.com/kundur98novel.html

[118] J. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright, "Convergence properties of the nelder-mead simplex method in low dimensions," *SIAM Journal of Optimization*, vol. 9, no. 1, pp. 112–147, 1998.

[119] R. L. Lagendijk and J. Biemond, "Block-adaptive image identification and restoration," in *Proceedings - ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 4, 1991, pp. 2497–2500.

[120] ——, "Basic methods for image restoration and identification," 1999, chapter 3.5, pages 125–139. Academic Press, 2000. [Online]. Available: http://citeseer.ist.psu.edu/lagendijk99basic.html

[121] R. L. Lagendijk, J. Biemond, and D. E. Boekee, "Regularized iterative image restoration with ringing reduction," *IEEE Transactions of Acoustics, Speech and Signal Processing*, vol. 36, no. 12, pp. 1874–1887, December 1988.

[122] ——, "Blur identification using the expectation-maximization algorithm," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 3, pp. 1397–1400, 1989.

[123] ——, "Identification and restoration of noisy blurred images using the expectation-maximization algorithm," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 7, pp. 1180–1191, Jul 1990.

[124] R. L. Lagendijk, A. Tekalp, and J. Biemond, "Maximum likelihood image and blur identification: a unifying approach," *Optical Engineering*, vol. 29, no. 5, pp. 422–435, 1990.

[125] R. G. Lane, "Blind deconvolution of speckle images," *Optical Society of America Journal A*, vol. 9, pp. 1508–1514, Sep 1992.

[126] R. G. Lane and R. H. T. Bates, "Automatic multidimensional deconvolution," *Journal of the Optical society of America-A*, vol. 4, no. 1, pp. 180–188, 1987.

[127] K. T. Lay and A. K. Katsaggelos, "Simultaneous identification and restoration of images using maximum likelihood estimation and the EM algorithm," in *Proc. 26th Annual Allerton Conf. On Commun., Control and Computing*, Monticello, IL, Sep 1988, pp. 661–662.

[128] K. Lay and A. K. Katsaggelos, "Image identification and restoration based on the expectation-maximization algorithm," *Optical Engineering*, vol. 29, no. 5, pp. 436–445, 1990.

[129] H.-C. Lee, "Review of image-blur models in a photographic system using the principles of optics," *Optical Engineering*, vol. 29, no. 5, pp. 405–421, May 1990.

[130] A. Levin, R. Fergus, F. Durand, and W. T. Freeman, "Image and depth from a conventional camera with a coded aperture," *ACM Transactions on Graphics (Proc. SIGGRAPH 2007)*, 2007. [Online]. Available: http://groups.csail.mit.edu/graphics/CodedAperture/CodedAperture-LevinEtAl-SIGGRAPH07.pdf

[131] C.-K. Liang, T.-H. Lin, B.-Y. Wong, C. Liu, and H. H. Chen, "Programmable aperture photography: Multiplexed light field acquisition," *ACM Transactions on Graphics (Proc. SIGGRAPH 2008)*, vol. 27, no. 3, 2008. [Online]. Available: http://mpac.ee.ntu.edu.tw/~chiakai/pap/

[132] A. Likas and N. Galatsanos, "A variational approach for Bayesian blind image deconvolution," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2222–2233, 2004.

[133] D. J. C. MacKay, "Probable networks and plausible predictions: Ůa review of practical Bayesian methods for supervised neural networks," *Network: Computation in Neural Systems*, vol. 6, no. 3, pp. 469–Ű505, 1995.

[134] J. Mairal, J. Mairal, G. Sapiro, and M. Elad, "Multiscale sparse image representationwith learned dictionaries," in *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, G. Sapiro, Ed., vol. 3, 2007, pp. III – 105–III – 108. [Online]. Available: http://www.cs.technion.ac.il/~elad/publications/conferences/2007/ICIP_2007_Submission.pdf

[135] J. Mait, R. Athale, and J. van der Gracht, "Evolutionary paths in imaging and recent trends," *Opt. Express*, vol. 11, no. 18, pp. 2093–2101, Sep. 2003. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-11-18-2093

[136] J. Marroquin, S. Mitter, and T. Poggio, "Probabilistic solution of ill-posed problems in computational vision," *Journal of the American Statistical Association*, vol. 82, no. 397, pp. 76–89, Mar 1987. [Online]. Available: http://links.jstor.org/sici?sici=0162-1459%28198703%2982%3A397%3C76%3APSOIPI%3E2.0.CO%3B2-D

[137] N. Mastronardi, P. Lemmerling, S. Van Huffel, A. Kalsi, and D. O'Leary, "Implementation of the regularized structured total least squares algorithms for blind image deblurring," *Linear Algebra and Its Applications*, vol. 391, no. 1-3 SPEC. ISS., pp. 203–221, 2004.

[138] B. McCallum, "Blind deconvolution by simulated annealing," *Optics Communications*, vol. 75, no. 2, pp. 101–105, 1990.

[139] V. Z. Mesarovic, N. P. Galatsanos, and A. K. Katsaggelos, "Regularized constrained total least–squares image restoration," *IEEE Trans. Image Process.*, vol. 4, no. 8, pp. 1096–1108, August 1995.

[140] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *The Journal of Chemical Physics*, vol. 21, no. 6, pp. 1087–1092, 1953. [Online]. Available: http://dx.doi.org/10.1063/1.1699114

[141] O. Michailovich and D. Adam, "A novel approach to the 2-D blind deconvolution problem in medical ultrasound," *IEEE Trans. Med. Imag.*, vol. 24, no. 1, pp. 86–104, 2005.

[142] O. V. Michailovich and D. R. Adam, "Deconvolution of medical images from microscopic to whole body images," in *Blind image deconvolution: Theory and Applications*, P. Campisi and K. Egiazarian, Eds. CRC, 2007, ch. 5.

[143] J. Miskin and D. J. C. MacKay, "Ensemble learning for blind image spearation and deconvolution," in *Advances in Independent Component Analysis*, M. Girolami, Ed. Springer-Verlag, 2000. [Online]. Available: http://www.inference.phy.cam.ac.uk/jwm1003/

[144] J. W. Miskin, "Ensemble learning for independent component analysis," Ph.D. dissertation, Astrophysics Group, Cavendish Laboratory, University of Cambridge, Dec 2000.

[145] N. Miura and N. Baba, "Extended-object reconstruction with sequential use of the iterative blind deconvolution method," *Optics Communications*, vol. 89, pp. 375–379, 1992.

[146] A. F. J. Moffat, "A theoretical investigation of focal stellar images in the photographic emulsion and application to photographic photometry," *Astronomy and Astrophysics*, vol. 3, pp. 455–461, 1969.

[147] R. Molina and B. D. Ripley, "Using spatial models as priors in astronomical image analysis," *Journal of Applied Statistics*, vol. 16, pp. 193–206, 1989.

[148] R. Molina, A. Katsaggelos, J. Mateos, and J. Abad, "Restoration of severely blurred high range images using compound models," *IEEE International Conference on Image Processing*, vol. 2, pp. 469–472, 1996. [Online]. Available: http://decsai.ugr.es/vip/files/conferences/96icip.pdf

[149] ——, "Compound Gauss-Markov random fields for astronomical image restoration," *Vistas in Astronomy (Special issue on "Vision Modeling and Information Coding")*, vol. 40, no. 4, pp. 539–546, 1997. [Online]. Available: http://decsai.ugr.es/vip/files/journals/95nice.pdf

[150] R. Molina, J. Mateos, A. Katsaggelos, C. Segall, and A. Hermoso, "Restoration of severely blurred high range images using stochastic and deterministic relaxation algorithms in compound Gauss-Markov random fields," *Pattern Recognition*, vol. 33, no. 4, pp. 555–571, 2000. [Online]. Available: http://ivpl.ece.northwestern.edu/Publications/Journals/PatternRecognition00.pdf

[151] R. Molina, J. Núñez, F. Cortijo, and J. Mateos, "Image restoration in astronomy: A Bayesian perspective," *IEEE Signal Processing Magazine*, vol. 18, no. 2, pp. 11–29, 2001. [Online]. Available: http://decsai.ugr.es/vip/files/journals/00spmagazine.pdf

[152] R. Molina, J. Mateos, A. Katsaggelos, and M. Vega, "Bayesian multichannel image restoration using compound Gauss-Markov random fields," *IEEE Transactions on Image Processing*, vol. 12, no. 12, pp. 1642–1654, 2003. [Online]. Available: http://decsai.ugr.es/vip/files/journals/TIP_2003_12_12_GMRF.pdf

[153] R. Molina, J. Mateos, and A. Katsaggelos, "Blind deconvolution using a variational approach to parameter, image, and blur estimation," *IEEE Trans Image Process*, vol. 15, no. 12, pp. 3715–3727, 2006. [Online]. Available: http://www.scopus.com/scopus/inward/record.url?eid=2-s2.0-33751368657&partnerID=40&rel=R5.5.0

[154] R. Molina, "On the hierarchical Bayesian approach to image restoration. applications to astronomical images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, pp. 1122–1128, 1994. [Online]. Available: http://decsai.ugr.es/vip/files/journals/92pami.pdf

[155] R. Molina, A. K. Katsaggelos, and J. Mateos, "Bayesian and regularization methods for hyperparameter estimation in image restoration," *IEEE Transactions on Image Processing*, vol. 8, no. 2, pp. 231–246, 1999. [Online]. Available: http://ivpl.ece.northwestern.edu/Publications/Journals/1999/IEEETransIP99.pdf

[156] J. Morgan, *Introduction to Geometrical and Physical Optics*. McGraw-Hill, 1953.

[157] J. G. Nagy and D. P. O'Leary, "Restoring images degraded by spatially variant blur," *SIAM Journal of Scientific Computing*, vol. 19, no. 4, pp. 1063–1082, 1998. [Online]. Available: http://www.cs.umd.edu/users/oleary/reprints/j45.pdf

[158] J. G. Nagy, R. J. Plemmons, and T. C. Torgersen, "Iterative image restoration using approximate inverse preconditioning," *IEEE Transactions on Image Processing*, vol. 5, no. 7, pp. 1151–1162, Jul 1996.

[159] R. Nakagaki and A. Katsaggelos, "A VQ-based blind image restoration algorithm," *Image Processing, IEEE Transactions on*, vol. 12, no. 9, pp. 1044–1053, Sep 2003.

[160] R. M. Neal, "Probabilistic inference using Markov chain Monte Carlo methods," Dept. of Computer Science, University of Toronto, University of Toronto, Tech. Rep. CRG-TR-93-1, 1993, available online at http://www.cs.toronto.edu/~radford/res-mcmc.html. [Online]. Available: citeseer.ist.psu.edu/neal93probabilistic.html

[161] R. M. Neal and G. E. Hinton, "A view of the em algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models*, M. Jordan, Ed. Cambridge, MA: MIT Press, 1998, pp. 355–368, also appears as Technical Report MSR-TR-95-06, Microsoft Research, March, 1995. An earlier version appears as Bayesian Networks for Data Mining, Data Mining and Knowledge Discovery, 1:79-119, 1997. [Online]. Available: http://research.microsoft.com/users/heckerma/

[162] M. K. Ng, R. H. Chan, and W.-C. Tang, "A fast algorithm for deblurring models with neumann boundary conditions," *SIAM J. Sci. Comput.*, vol. 21, no. 3, pp. 851–866, Jan. 1999.

[163] M. K. Ng, R. J. Plemmons, and F. Pimentel, "A new approach to constrained total least squares image restoration," *Linear Algebra and Its Applications*, vol. 316, no. 1-3, pp. 237–258, 2000.

[164] M. Ng, R. Plemmons, and S. Qiao, "Regularization of RIF blind image deconvolution," *Image Processing, IEEE Transactions on*, vol. 9, no. 6, pp. 1130–1134, 2000.

[165] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," Stanford University, Tech. Rep. CSTR 2005-02, April 2005. [Online]. Available: http://graphics.stanford.edu/papers/lfcamera/

[166] K. Nishi and S. Ando, "Blind superresolving image recovery from blur-invariant edges," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 5, pp. 85–88, 1994.

[167] J. J. Ó Ruanaidh and W. Fitzgerald, *Numerical Bayesian Methods Applied to Signal Processing*, 1st ed., ser. Springer Series in Statistics and Computing. New York: Springer, 1996, iSBN 0-387-94629-2.

[168] C. A. Ong and J. Chambers, "An enhanced NAS-RIF algorithm for blind image deconvolution," *Image Processing, IEEE Transactions on*, vol. 8, no. 7, pp. 988–992, 1999.

[169] M. K. Ozkan, A. Tekalp, and M. Sezan, "Identification of a class of space-variant image blurs," *Proceedings of SPIE – The International Society for Optical Engineering*, vol. 1452, pp. 146–156, 1991.

[170] H.-T. Pai and A. C. Bovik, "On eigenstructure-based direct multichannel blind image restoration," *Image Processing, IEEE Transactions on*, vol. 10, no. 10, pp. 1434–1446, 2001.

[171] ——, "Exact multichannel blind image restoration," *IEEE Signal Processing Letters*, vol. 4, no. 8, pp. 217–220, 1997.

[172] H.-T. Pai, J. Havlicek, and A. Bovik, "Generically sufficient conditions for exact multichannel blind image restoration," in *Acoustics, Speech, and Signal Processing, 1998. ICASSP '98. Proceedings of the 1998 IEEE International Conference on*, vol. 5, 1998, pp. 2861–2864.

[173] K. Panchapakesan, D. G. Sheppard, M. W. Marcellin, and B. R. Hunt, "Blur identification from vector quantizer encoder distortion," *IEEE Trans. Image Process.*, vol. 10, pp. 465–470, Mar 2001.

[174] S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: A technical overview," *IEEE Signal Processing Magazine*, vol. 20, no. 3, pp. 21–36, 2003.

[175] L. C. Pickup, D. P. Capel, S. J. Roberts, and A. Zisserman, "Overcoming registration uncertainty in image super-resolution: Maximize or marginalize?" *EURASIP Journal on Advances in Signal Processing*, vol. 2007, pp. Article ID 23 565, 14 pages, 2007.

[176] ——, "Bayesian methods for image super-resolution," *The Computer Journal*, p. bxm091, Oct. 2007. [Online]. Available: http://comjnl.oxfordjournals.org/cgi/content/abstract/bxm091v1

[177] J. Portilla, V. Strela, M. Wainwright, and E. Simoncelli, "Image denoising using scale mixtures of gaussians in the wavelet domain," *Image Processing, IEEE Transactions on*, vol. 12, no. 11, pp. 1338–1351, 2003.

[178] W. K. Pratt, "Vector space formulation of two-dimensional signal processing operations," *Computer Graphics and Image Processing*, vol. 4, no. 3, pp. 1–24, Sep 1975.

[179] ——, *Digital Image Processing*. John Wiley & Sons, 1978.

[180] H. Raiffa and R. Schlaifer, *Applied Statistical Decision Theory*. Division of Research, Graduate School of Business, Administration, Harvard University, Boston, 1961.

[181] A. Rajagopalan and S. Chaudhuri, "A recursive algorithm for maximum likelihood-based identification of blur from multiple observations," *IEEE Transactions on Image Processing*, vol. 7, no. 7, pp. 1075–1079, 1998. [Online]. Available: http://dx.doi.org/10.1109/83.701169

[182] ——, "MRF model-based identification of shift-variant point spread function for a class of imaging systems," *Signal Processing*, vol. 76, no. 3, pp. 285–299, 1999.

[183] R. Raskar, A. Agrawal, and J. Tumblin, "Coded exposure photography: motion deblurring using fluttered shutter," *ACM Trans. Graph. (Proc. SIGGRAPH 2006)*, vol. 25, no. 3, pp. 795–804, 2006. [Online]. Available: http://www.cfar.umd.edu/~aagrawal/sig06/CodedExposureLowres.pdf

[184] A. Rav-Acha and S. Peleg, "Two motion-blurred images are better than one," *Pattern Recognition Letters*, vol. 26, no. 3, pp. 311–317, 2005.

[185] S. F. Ray, *Applied photographic optics*. Focal Press, 2002.

[186] S. J. Reeves and R. M. Mersereau, "Optimal estimation of the regularization parameter and stabilizing functional for regularized image restoration," *Optical Engineering*, vol. 29, no. 5, pp. 446–454, 1990.

[187] ——, "Blur identification by the method of generalized cross-validation," *IEEE Transactions on Image Processing*, vol. 1, no. 3, pp. 301–311, 1992.

[188] W. H. Richardson, "Bayesian-based iterative method of image restoration," *Optical Society of America Journal A*, vol. 62, no. 1, pp. 55–59, 1972.

[189] B. D. Ripley, *Spatial Statistics*. John Wiley, 1981, pp. 88–90.

[190] S. Roth and M. Black, "Fields of experts: a framework for learning image priors," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2, 2005, pp. 860–867 vol. 2.

[191] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1-4, pp. 259–268, Nov. 1992. [Online]. Available: http://www.sciencedirect.com/science/article/B6TVK-46JYGGT-GH/2/f85f57585ec9af00a1a088d0b8e6d452

[192] A. E. Savakis and H. J. Trussell, "On the accuracy of PSF representation in image restoration," *IEEE Transactions on Image Processing*, vol. 2, no. 2, pp. 252–259, 1993.

[193] R. R. Schultz and R. L. Stevenson, "Bayesian approach to image expansion for improved definition," *IEEE Transactions on Image Processing*, vol. 3, no. 3, pp. 233–242, 1994.

[194] ——, "Extraction of high-resolution frames from video sequences," *IEEE Transactions on Image Processing*, vol. 5, no. 6, pp. 996–1011, 1996.

[195] T. J. Schulz, "Multiframe blind deconvolution of astronomical images," *Journal of the Optical society of America-A*, vol. 10, pp. 1064–1073, 1993.

[196] C. Segall, R. Molina, and A. Katsaggelos, "High-resolution images from low-resolution compressed video," *IEEE Signal Processing Magazine*, vol. 20, no. 3, pp. 37–48, 2003.

[197] S. Serra-Capizzano, "A note on antireflective boundary conditions and fast deblurring models," *Siam J. Sci. Comput.*, vol. 25, no. 4, pp. 1307–1325, 2003. [Online]. Available: http://www.scopus.com/scopus/inward/record.url?eid=2-s2.0-4043054296&partner=40&rel=R5.0.4

[198] M. Sezan, G. Pavlovic, A. Tekalp, and A. Erdem, "On modeling the focus blur in image restoration," in *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*, 1991, pp. 2485–2488 vol.4.

[199] M. Sezan and A. Tekalp, "Survey of recent developments in digital image restoration," *Optical Engineering*, vol. 29, no. 5, pp. 393–404, 1990.

[200] J. R. Shewchuk, "An introduction to the conjugate gradient method without the agonizing pain," August 1994. [Online]. Available: http://www.cs.cmu.edu/~jrs/jrspapers.html#cg

[201] V. Smídl and A. Quinn, *The Variational Bayes Method in Signal Processing*. Springer, 2005.

[202] M. M. Sondhi, "Image restoration: The removal of spatially invariant degradations," *Proceedings of the IEEE*, vol. 60, no. 7, pp. 842–853, Jul 1972.

[203] M. Sorel and J. Flusser, "Space-variant restoration of images degraded by camera motion blur," *Image Processing, IEEE Transactions on*, vol. 17, no. 2, pp. 105–116, 2008.

[204] F. Šroubek, G. Cristóbal, and J. Flusser, "A unified approach to superresolution and multichannel blind deconvolution," *Image Processing, IEEE Transactions on*, vol. 16, no. 9, pp. 2322–2332, 2007.

[205] F. Šroubek and J. Flusser, "Multichannel blind iterative image restoration," *Image Processing, IEEE Transactions on*, vol. 12, no. 9, pp. 1094–1106, 2003. [Online]. Available: http://www.utia.cas.cz/library/prace/20030104.pdf

[206] ——, "Multichannel blind deconvolution of spatially misaligned images," *IEEE Transactions on Image Processing*, vol. 7, pp. 45–53, Jul 2005. [Online]. Available: http://www.utia.cas.cz/sroubekf/

[207] ——, "An overview of multichannel image restoration techniques," in *Week of Doctoral Students*, J. Safránková, Ed. Praha: Matfyzpress, 1999, pp. 580–585. [Online]. Available: http://www.utia.cas.cz/library/prace/990225.ps

[208] T. G. Stockham, Jr., T. M. Cannon, and R. B. Ingebretsen, "Blind deconvolution through digital signal processing." *Proceedings of the IEEE*, vol. 63, no. 4, pp. 678–692, 1975.

[209] M. Subbarao, "On the depth information in the point spread function of a defocused optical system," Computer Vision Laboratory, Department of Electrical Engineering State University of New York, Stony Brook, NY, 11794-2350, Tech. Rep. 90.02.07, Feb 1990. [Online]. Available: citeseer.ist.psu.edu/subbarao99depth.html

[210] A. M. Tekalp, H. Kaufman, and J. W. Woods, "Identification of image and blur parameters for the restoration of noncausal blurs," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 963–972, Aug 1986.

[211] A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-Posed Problems*. Washington D.C.: John Wiley & Sons, August 1977.

[212] M. E. Tipping and C. M. Bishop, "Bayesian image super-resolution," in *Advances in Neural Information Processing Systems*, S. Becker, S. Thrun, and K. Obermeyer, Eds., vol. 15, 2002, pp. 1303–1310. [Online]. Available: http://research.microsoft.com/~cmbishop/publications_abs.htm#Bishop:superres02

[213] B. C. Tom, K.-T. Lay, and A. K. Katsaggelos, "Multi-channel image identification and restoration using the expectation-maximization algorithm," *Optical Engineering*, vol. 35, no. 1, pp. 241–254, 1996. [Online]. Available: http://link.aip.org/link/?JOE/35/241/1

[214] P. T. Troughton, "Simulation methods for linear and nonlinear time series models with application to distorted audio signals," Ph.D. dissertation, University of Cambridge, Department of Engineering, Jun 1999.

[215] H. Trussell, "Relationship between image restoration by the maximum a posteriori method and a maximum entropy method." *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-28, no. 1, pp. 114–117, 1980.

[216] F. Tsumuraya, N. Miura, and N. Baba, "Iterative blind deconvolution method using Lucy's algorithm," *Astronomy and Astrophysics*, vol. 282, no. 2, pp. 699–708, 1994.

[217] D. Tzikas, A. Likas, and N. Galatsanos, "Variational bayesian blind image deconvolution with student-t priors," in *IEEE International Conference on Image Processing (ICIP)*, 2007.

[218] G. Van Kempen, L. Van Vliet, P. Verveer, and H. Van Der Voort, "A quantitative comparison of image restoration methods for confocal microscopy," *Journal of Microscopy*, vol. 185, no. 3, pp. 354–365, 1997. [Online]. Available: http://www.ph.tn.tudelft.nl/People/lucas/publications/1997/JoM97GKLVea/JoM97GKLVea.pdf

[219] A. Veeraraghavan, R. Raskar, A. K. Agrawal, A. Mohan, and J. Tumblin, "Dappled photography: mask enhanced cameras for heterodyned light fields and coded aperture refocusing," *ACM Trans. Graph. (Proc SIGGRAPH 2007)*, vol. 26, no. 3, p. 69, 2007. [Online]. Available: http://www.merl.com/people/agrawal/sig07/index.html

[220] C. Vogel and M. Oman, "Iterative methods for total variation denoising," *SIAM Journal of Scientific Computing*, vol. 17, no. 1, pp. 227–238, 1996.

[221] M. Wakin, J. Laska, M. Duarte, D. Baron, S. Sarvotham, D. Takhar, K. Kelly, and R. Baraniuk, "An architecture for compressive imaging," in *Image Processing, 2006 IEEE International Conference on*, 2006, pp. 1273–1276.

[222] Z. Wang and F. Qi, "Analysis of multiframe super-resolution reconstruction for image anti-aliasing and deblurring," *Image and Vision Computing*, vol. 23, no. 4, pp. 393–404, Apr 2005.

[223] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity." *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, vol. 13, no. 4, pp. 600–612, 2004.

[224] J. Weickert, "A review of nonlinear diffusion filtering," in *SCALE-SPACE '97: Proceedings of the First International Conference on Scale-Space Theory in Computer Vision*. London, UK: Springer-Verlag, 1997, pp. 3–28. [Online]. Available: http://portal.acm.org/citation.cfm?id=715066

[225] R. A. Wiggins, "Minimum entropy deconvolution." *Geoexploration*, vol. 16, no. 1-2, pp. 21–35, 1978. [Online]. Available: http://www.scopus.com/scopus/inward/record.url?eid=2-s2.0-0017961352&partner=40&rel=R4.5.0

[226] C. S. Won and R. M. Gray, *Stochastic Image Processing*, ser. Information Technology: Transmission, Processing, and Storage. Kluwer Academic / Plenum Publishers, 2004.

[227] J. W. Woods and V. K. Ingle, "Kalman filtering in two dimensions: further results." *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-29, no. 2, pp. 188–197, 1981.

[228] J. W. Woods and C. H. Radewan, "Kalman filtering in two dimensions." *IEEE Transactions on Information Theory*, vol. IT-23, no. 4, pp. 473–482, 1977.

[229] J. Woods, "Two-dimensional discrete markovian fields," *IEEE Transactions on Information Theory*, vol. IT18, no. 2, pp. 232–240, Mar 1972.

[230] ——, "Markov image modeling," *IEEE Transactions on Automatic Control*, vol. AC-23, no. 5, pp. 846–50, Oct 1978. [Online]. Available: http://dx.doi.org/10.1109/TAC.1978.1101866

[231] H.-S. Wu, "Minimum entropy deconvolution for restoration of blurred two-tone images," *Electronics Letters*, vol. 26, no. 15, pp. 1183–1184, 1990. [Online]. Available: http://www.scopus.com/scopus/inward/record.url?eid=2-s2.0-0025463254&partner=40&rel=R4.5.0

[232] Y. Yang, N. P. Galatsanos, and H. Stark, "Projection-based blind deconvolution," *Journal of the Optical Society of America A: Optics and Image Science, and Vision*, vol. 11, no. 9, pp. 2401–2409, 1994.

[233] K.-H. Yap, L. Guan, and W. Liu, "A recursive soft-decision approach to blind image deconvolution," *IEEE Transactions on Signal Processing*, vol. 51, no. 2, pp. 515–526, 2003.

[234] Y.-L. You and M. Kaveh, "Ringing reduction in image restoration by orientation-selective regularization," *IEEE Signal Processing Letters*, vol. 3, no. 2, pp. 29–31, 1996.

[235] ——, "A regularization approach to joint blur identification and image restoration," *Image Processing, IEEE Transactions on*, vol. 5, no. 3, pp. 416–428, Mar 1996.

[236] ——, "Blind image restoration by anisotropic regularization," *Image Processing, IEEE Transactions on*, vol. 8, no. 3, pp. 396–407, 1999.

[237] Y.-L. You, W. Xu, A. Tannenbaum, and M. Kaveh, "Behavioral analysis of anisotropic diffusion in image processing," *IEEE Transactions on Image Processing*, vol. 5, no. 11, pp. 1539–1553, 1996.

[238] J. Zhang, "The mean field theory in EM procedures for Markov random fields," *IEEE Transactions on Signal Processing*, vol. 40, no. 10, pp. 2570–2583, 1992.

[239] ——, "The mean field theory in EM procedures for blind Markov random field image restoration," *Image Processing, IEEE Transactions on*, vol. 2, no. 1, pp. 27–40, 1993.